NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

OPTIMIZING CLASSIFICATION IN INTELLIGENCE PROCESSING

by

Yinon Costica

December 2010

Thesis Co-Advisors: Moshe Kress
Roberto Szechtman
Second Reader: Patricia Jacobs

Approved for public release; distribution is unlimited

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** December 2010	**3. REPORT TYPE AND DATES COVERED** Master's Thesis

4. TITLE AND SUBTITLE
Optimizing Classification in Intelligence Processing

5. FUNDING NUMBERS

6. AUTHOR(S) Yinon Costica

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Postgraduate School
Monterey, CA 93943-5000

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES)
N/A

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (maximum 200 words)

The intelligence making process, often described as the intelligence cycle, consists of phases. Congestion may be experienced in phases that require time consuming tasks such as translation, processing and analysis. To ameliorate the performance of those time-consuming phases, a preliminary classification of intelligence items regarding their relevance and value to an intelligence request is performed. This classification is subject to false positive and false negative errors, where an item is classified as positive if it is relevant and provides valuable information to an intelligence request, and negative otherwise. The tradeoff between both types of errors, represented visually by the Receiver Operating Characteristic curve, depends on the training and capabilities of the classifiers as well as the classification test performed on each item and the decision rule that separates between positives and negatives.

An important question that arises is how to best tune the classification process such that both accuracy of the classification and its timeliness are adequately addressed. An analytic answer is presented via a novel optimization model based on a tandem queue model.

This thesis provides decision makers in the intelligence community with measures of effectiveness and decision support tools for enhancing the effectiveness of the classification process in a given intelligence operations scenario. In addition to the analytic study, numerical results are presented to obtain quantitative insights via sensitivity analysis of input parameters.

14. SUBJECT TERMS
ROC Curve, Intelligence Cycle, Queuing Model, Intelligence Processing, Intelligence Analysis, Binary Classification

15. NUMBER OF PAGES
75

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** UU

THIS PAGE INTENTIONALLY LEFT BLANK

OPTIMIZING CLASSIFICATION IN INTELLIGENCE PROCESSING

Yinon Costica
Captain, Israel Defense Forces
B.Sc. Computer Sciences, Mathematics and Physics,
The Hebrew University of Jerusalem, 2004

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
December 2010

Author: Yinon Costica

Approved by: Moshe Kress
Thesis Co-Advisor

Roberto Szechtman
Thesis Co-Advisor

Patricia Jacobs
Second Reader

Robert F. Dell
Chairman, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

The intelligence making process, often described as the intelligence cycle, consists of phases. Congestion may be experienced in phases that require time-consuming tasks such as translation, processing and analysis. To ameliorate the performance of those time-consuming phases, a preliminary classification of intelligence items regarding their relevance and value to an intelligence request is performed. This classification is subject to false positive and false negative errors, where an item is classified as positive if it is relevant and provides valuable information to an intelligence request, and negative otherwise. The tradeoff between both types of errors, represented visually by the Receiver Operating Characteristic curve, depends on the training and capabilities of the classifiers as well as the classification test performed on each item and the decision rule that separates between positives and negatives.

An important question that arises is how to best tune the classification process such that both accuracy of the classification and its timeliness are adequately addressed. An analytic answer is presented via a novel optimization model based on a tandem queue model.

This thesis provides decision makers in the intelligence community with measures of effectiveness and decision support tools for enhancing the effectiveness of the classification process in a given intelligence operations scenario. In addition to the analytic study, numerical results are presented to obtain quantitative insights via sensitivity analysis of input parameters.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

ACC	Classification Accuracy
AUC	Area Under the ROC Curve
CI	Competitive Intelligence
COMINT	Communications Intelligence
DoD	Department of Defense
ELINT	Electronic Intelligence
FN, FNR	False Negative, False Negative Rate
FP, FPR	False Positive, False Positive Rate
GAMS	General Algebraic Modeling System
HUMINT	Human Intelligence
IMINT	Imagery Intelligence
IR	Intelligence Requirements
ISR	Intelligence, Surveillance and Reconnaissance
KKT	Karush-Kuhn-Tucker
MOE	Measure of Effectiveness
N	Negative
OR	Operations Research
OSINT	Open Source Intelligence
P	Positive
PR	Precision-Recall
ROC	Receiver Operating Characteristic
SIGINT	Signal Intelligence

THIS PAGE INTENTIONALLY LEFT BLANK

EXECUTIVE SUMMARY

Intelligence provides leaders with information and knowledge to support decision making. The process of producing intelligence is commonly described as a cycle of phases that begins with the issuing of an information request and ends with the dissemination of a coherent assessment to the relevant consumers. The new information leads to an update of the information requests and hence the cyclic nature of the process.

Today's vast usage of communication results in a glut of information, which may create bottlenecks in phases that require close attention to each information item such as translation, processing and analysis. Moreover, as the timeliness of the information becomes more crucial, the bottleneck's effect becomes more critical.

To ameliorate the performance of those time-consuming phases, a preliminary classification of items as to their relevance to an information request is performed. Nevertheless, this binary classification process requires additional resources such as personnel and time, and it is subject to false positive and false negative errors, where an item is classified as positive if it is relevant and provides additional information to an intelligence request and negative otherwise. The tradeoff between both types of errors is represented by a functional relationship called the Receiver Operating Characteristic (ROC) curve.

The quality of the classification, as manifested by the prevalence of both types of errors on the ROC curve, depends on several key parameters. Some are strategic parameters of the system, such as the training of the classifiers and the overhead costs involved with the system, and some are tactical parameters that can be adjusted within the operation of a given classification system, such as the time spent on the classification of each item and the decision mechanism that is used to determine the nature of the item according to the classification rule.

An important issue that arises concerns the optimal tuning of the tactical parameters of the classification process so that both the accuracy of the classification and the timeliness of the resulting intelligence product fall within certain desired ranges. We

consider models in which increased classification skill is associated with increase in the mean classification service time. The model and analysis presented in this thesis are a first modeling step towards balancing investments in the intelligence cycle, a key operations research related issue that was emphasized by the Defense Science Board's Advisory Group on Defense Intelligence, in a report on “Operations Research Applications for Intelligence, Surveillance and Reconnaissance (ISR)” in 2009.

In this thesis we make the following contributions: first, we provide decision makers in the intelligence community with measures of effectiveness (MOEs) to assess the classification process in a given intelligence operations scenario. An item is considered a true positive if a well-trained analyst, having enough time and resources, would classify it as positive. Two MOEs are suggested: the first measures the performance of the classification in terms of the achieved true positive rate, and the second measures its cost-effectiveness, where cost is assumed to be driven by the time spent on each item and the effectiveness is measured by the number of positive items produced. False positives are not directly penalized; however, the analysis time required to process them reduces the efficiency of the system. An analytic answer for the aforementioned tradeoff between classification accuracy and timeliness is obtained via an optimization model based on a tandem queuing model for classifying intelligence items in the presence of limited resources and time constraints; the model assumes that the contribution of an item does not change while it is waiting to be processed. The optimization model adjusts the values of the classification process tactical parameters in order to maximize the first performance MOE, namely the achieved true positive rate.

The effectiveness of the classification, as manifested by the true positive rate achieved at optimality, is measured with respect to the true positive rate of the system when no classification is implemented at all. This measure allows us not only to quantify the benefit of the classification under a given scenario, but also to compare the added value among different scenarios, allowing thumb rules for better classification resource allocation.

The main parameters and relationships that affect the performance of the classification process are identified and used in developing the model. Based on the optimal results, we obtain measures of effectiveness for decision makers to compare the performance of the classification in different scenarios. In addition to an analytic study of the model, which results in qualitative insight, we also discuss numerical results to obtain quantitative insights.

Using the implemented model, different intelligence operations scenarios are studied and compared. We consider two scenarios of timeliness: a tactical scenario such as tactical engagements on the battlefield in which intelligence information is needed quickly (e.g. “ticking time bomb” scenario), and a strategic scenario such as long term armament transactions. Three scenarios are considered with respect to the source quality, which is defined as the fraction of true positive items among all items in the source: low, medium and high value sources.

We have shown that for the implemented model, the larger the fraction of items in the source that are true positives (higher quality source) the less beneficial it is, in both measures of effectiveness, to implement the classification. In addition, for low quality sources (i.e., fewer true positives) classification improves the true positive rate for tactical scenarios more than for strategic ones. For high quality sources, the opposite applies. In addition, a cost-effectiveness study of the relationship between the classifier cost and the classification capability limit is developed; this relationship can be used to compare different classifiers. Specifically it is shown that for a high-quality source, it is more cost-effective to allow the analysts to directly use items from the source without pre-classification, despite the limited resources. Given the cost of the classification, the breakeven source quality in which both alternatives, namely with and without pre-classification, bear the same cost can be estimated.

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

I wish to express my deep appreciation to all those who supported me in this endeavor: to my family and friends who felt so close despite the distance, and to those who made California feel like home.

A special gratitude to my advisors, Moshe Kress and Roberto Szechtman, for their patient and enlightening guidance, motivation and, most importantly, friendship.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

A. INTRODUCTION

Intelligence provides leaders with information and knowledge to support decision making. The process of producing intelligence is commonly described as a cycle of phases that begins with the issuing of an information request and ends with the dissemination of a coherent assessment to the relevant consumers. The new information leads to an update of the information requests and hence the cyclic nature of the process.

Today's vast usage of communication results in a glut of information, which may create bottlenecks in phases that require close attention to each information item such as translation, processing and analysis. Moreover, as the timeliness of the information becomes more crucial, the bottleneck's effect becomes more critical.

To ameliorate the performance of those time-consuming phases, a preliminary classification of items is performed to distinguish positive items from negative ones. This involves an item being classified as positive if it is relevant and provides additional information to an intelligence request and negative otherwise. A true positive is an item that would be classified by the analyst as positive. Nevertheless, this binary classification process requires additional resources such as personnel and time, and it is subject to false positive and false negative errors. The tradeoff between both types of errors is represented by a functional relationship called the Receiver Operating Characteristic (ROC) curve.

The quality of the classification, as manifested by the prevalence of both types of error on the ROC curve, depends on several key parameters. Some are strategic parameters of the system, such as the training of the classifiers and the overhead costs involved with the system, and some are tactical parameters that can be adjusted within the operation of a given classification system, such as the time spent on the classification of each item and the decision mechanism that is used to determine the nature of the item according to the classification rule.

B. MOTIVATION AND RESEARCH FOCUS

The motivation of this research is two-fold. First, this research provides decision makers in the intelligence community with measures of effectiveness (MOEs) to assess the classification process in a given intelligence operations scenario, where a desired item in terms of relevance is called a positive or if otherwise, negative. Two MOEs are suggested: the first measures the performance of the classification in terms of the achieved false negative rate, and the second measures its cost-effectiveness, where cost is assumed to be driven by the time spent on each item and the effectiveness is measured by the number of correctly identified positive items. False positives are not directly penalized; however, the analysis time required in order to process them reduces the efficiency of the system.

The second motivation is to optimize the tactical parameters of the classification process with respect to the first performance MOE, and study the effect of input parameters, such as inflow rate, as well as the strategic parameters on both MOEs via sensitivity analysis.

In this thesis, we develop a queuing model, embedded in an optimization model, which provides an analytical solution for the question of how to best tune the classification process such that both accuracy of the classification, as well as its timeliness, are adequately addressed. The model is implemented using the General Algebraic Modeling System (GAMS) and solved using the CONOPT3 non-linear programming solver (Brooke et al., 1998; Drud, 2005).

C. CONTRIBUTIONS OF THIS WORK

The main contribution of this thesis is a novel optimization model for classifying intelligence items in the presence of limited resources and time constraints. The main parameters and relationships that affect the performance of the classification process are identified and used in developing the model. Based on the optimal results, we obtain measures of effectiveness for decision makers to compare the performance of the classification in different scenarios. In addition to an analytic study of the model, which results in qualitative insight, we also implement the model numerically and obtain quantitative insights.

D. STRUCTURE OF THE THESIS AND CHAPTER OUTLINE

This thesis has five chapters. Following Chapter I (Introduction), Chapter II provides background information on the intelligence process and, specifically, on the problem addressed in this thesis, which is optimizing a binary classification process. In Chapter III, the operational setting is presented and the model is formulated and discussed. In Chapter IV, we present numerical results from the model and provide a brief sensitivity analysis of the input variables. Chapter V summarizes the research, presents the main findings and insights, and discusses potential future work in the area.

THIS PAGE INTENTIONALLY LEFT BLANK

II. BACKGROUND

A. INTELLIGENCE AS A PRODUCT

Intelligence is defined by the U.S. Department of Defense (DoD) as "the product resulting from the collection, processing, integration, evaluation, analysis, and interpretation of available information concerning foreign nations, hostile or potentially hostile forces or elements, or areas of actual or potential operations."[1] Besides the use of intelligence as an indispensible tool to support a national leader's decision making process, *competitive intelligence* (CI) has emerged in recent decades as an environment meant to provide a competitive edge for a privately owned organization (Khaner, 1998).

Thus, intelligence is a product of a process termed the *intelligence process*, defined by the DoD as "the process by which information is converted into intelligence and made available to users." The root of the intelligence process lies in the need for information by decision makers at every level.

The intelligence process is most commonly described as a feedback process called the *intelligence cycle*, which is a continuous investigation that allows decision makers to collect relevant information for supporting informed decisions. The *intelligence cycle* consists of five key elements[2,3] (see Figure 1): (1) Planning and Direction (2) Collection (3) Processing and Exploitation (4) Analysis and Production (5) Dissemination.

[1] DoD Dictionary of Military Terms, "Intelligence," http://www.dtic.mil/doctrine/dod_dictionary/data/i/4850.html Intelligence.

[2] Central Intelligence Agency, "Factbook on Intelligence: The Intelligence Cycle," http://www.fas.org/irp/cia/product/fact97/intcycle.htm.

[3] FBI, Intelligence Cycle, http://www.fbi.gov/about-us/intelligence/intelligence-cycle.

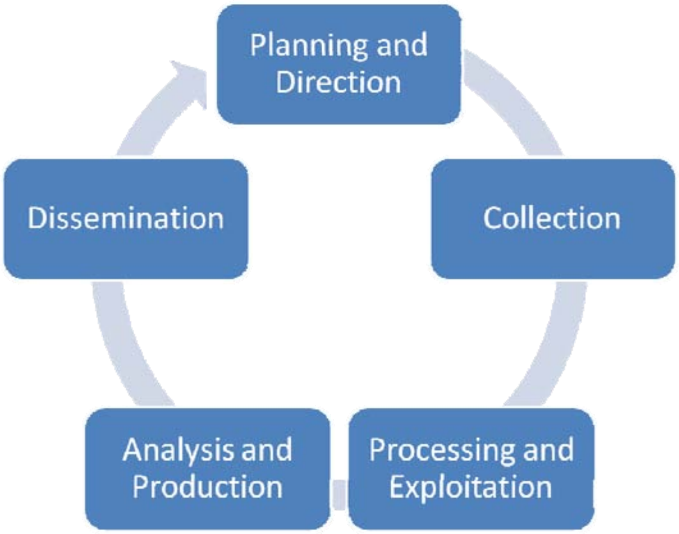

Figure 1. The intelligence cycle

The intelligence cycle is merely a model describing the outline of the intelligence production process and, like any other model, is a simplification and abstraction of a much more complex process. Despite several arguments that have been raised against the oversimplification of the intelligence process as the intelligence cycle (Richelson, 1999; Hulnick, 2006, it is useful for exploring tradeoffs and interactions among its components.

Miller et al. (2004) constructed an aggregated simulation model according to the aforementioned intelligence cycle that allows comparisons between different structures of the intelligence process. The comparison is based on four intuitive measures of performance: quality, quantity, timeliness, and satisfaction of information needs. Bose (2008) lists six measures of effectiveness for the intelligence product in competitive intelligence environment: accuracy, objectivity, usability, relevance, readiness, and timeliness. Nevertheless, each phase of the intelligence cycle may require adapted measures that refer to its specific role in the cycle.

1. Planning and Direction

The intelligence cycle is initiated with the identification of information needs, which are called *intelligence requirements* (IR). An intelligence agency is usually required to produce intelligence that serves decision making with respect to a list of IRs issued by possibly different users. The planning and direction phase identifies these IRs and prioritize them so subsequent phases will adjust their operation accordingly. The planning and direction responds to the outcomes at the end of the intelligence cycle as well, since the delivered intelligence generates new requirements and may result in re-prioritization of the existing IRs.

2. Collection

The collection of intelligence incorporates a variety of means to gather raw information from which subsequent phases produce finished products. Many different intelligence collection disciplines exist (Johnson & Wirtz, 2004), most notably: (1) Human Intelligence (HUMINT) collects information from persons such as agents and defectors, (2) Signal Intelligence (SIGINT) collects information by intercepting signals between people such as communication lines (known as communications intelligence, COMINT) and electronic emissions not intended for communications (known as electronic intelligence, ELINT), (3) Imagery Intelligence (IMINT) collects information from imaging systems such as satellite images and aerial photography and (4) Open Source Intelligence (OSINT) which aims at harvesting open sources such as TV, radio and newspapers for information. This source of information has been given new life with the spread of the internet and the proliferation of information in other broadcasting media (Hulnick, 2006).

In order to collect the raw information from available sources, an *intelligence collection plan* is generated. The collection plan should provide enough raw information for subsequent intelligence cycle elements so that the IRs issued by the planners are adequately satisfied. The collection plan follows the IRs' priorities because usually the collection resources are scarce compared to the spectrum of requirements and potential collection efforts and sources.

3. Processing and Exploitation

The processing phase is designated to transform the raw information that was collected into products that may be used in the analysis phase. The nature of the collected raw material dictates the type of operations to be included in the processing effort, as well as the required analysis capabilities. Common processing operations include data reduction, noise reduction, decryption, language translations, context clarification and more. Processing also includes the loading of the collected data into easily accessible databases where it can be exploited for use by the analysts later on.

The processing phase is the first contact with the raw information after it has been collected, and it is usually performed by those who have the background to understand the environmental and operational context in which it was collected. For example, when processing a raw IMINT product, technical parameters such as the location from which the image was taken, the time of the day, etc., can enhance the information content of the raw product in order to give the analyst a richer context. We call the people and technologies used during the processor phase the processors.

In many scenarios, processing resources can become a bottleneck, e.g. limited number of translators, limited computing resources, etc. When such limitations are present, a preliminary classification, in which some collected items are filtered out, must be done in order to reduce the flow of information and thus affect the processing throughput. Even tasks that simply require formatting and context enrichment of an item will create a bottleneck when the amount of collected information increases. This is particularly the case when it comes to SIGINT, which deals with a glut of information requiring relatively significant processing effort that includes translation and context enrichment.

4. Analysis and Production

Johnston (2005) defines intelligence analysis as “the application of individual and collective cognitive methods to weigh data and test hypotheses within a secret socio-cultural context.”[4]

The job of the analyst, as described by Hulnick (2006), is to evaluate the relevancy of the processed intelligence and put it in perspective with respect to current

[4] Analytic Culture in the U.S. Intelligence Community, Chapter One, https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/books-and-monographs/analytic-culture-in-the-u-s-intelligence-community/chapter_1.htm.

assessments. The FBI[5] includes under the analysis phase the integration, evaluation and analysis of the available data, and the subsequent preparation of the final intelligence product.

a. The Foraging and Sense Making Loops

Card and Pirolli (2005) provided an empirical descriptive study of the intelligence analysis using a cognitive task analysis. In their study, the analysis process is separated into two loops: (1) the *foraging loop* in which valuable information is culled and becomes *evidence* (Pirolli & Card, 1999), and (2) the *sense-making loop* (Russell et al., 1993) in which an integrated story that explains and presents the gathered evidence is iteratively developed. The process is described in Figure 2.

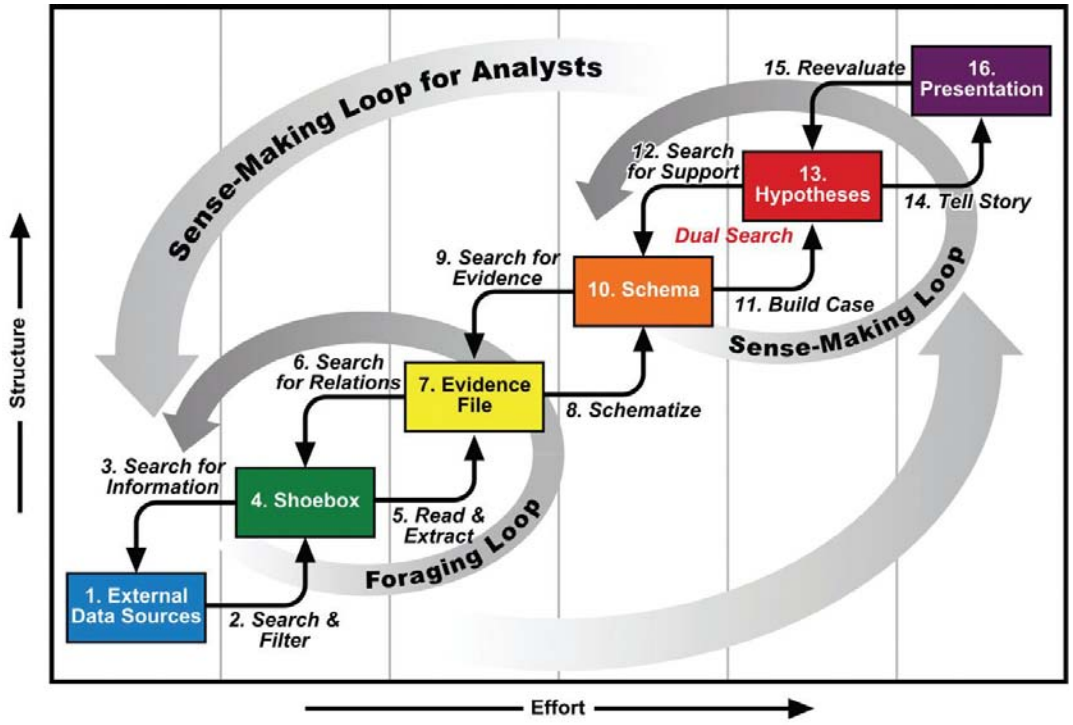

Figure 2. Nominal sense-making loop for some types of intelligence analysts[6]

[5] FBI, Directorate of Intelligence, Intelligence Cycle, http://www.fbi.gov/about-us/intelligence/intelligence-cycle.

[6] National Visualization and Analytics Center, Illuminating the Path: The Research and Development Agenda for Visual Analytics, p 44. http://nvac.pnl.gov/docs/RD_Agenda_NVAC_chapter2.pdf

An important difference between the two loops in our context is that the *foraging loop*, which follows the processing phase, deals with each item independently until it becomes evidence. On the other hand, the *sense-making loop* deals with the integration of a collection of evidence files and their presentation, rather than with the independent evidence files. Therefore, measuring and referring to intelligence material in terms of items is relevant up to the sense-making loop.

The foraging loop begins by searching and filtering the valuable items from external, usually vast, repositories. The relevant items are gathered in a virtual holder called the *shoebox*. The shoebox items are read and used to create evidence files, which are the building blocks of the sense-making loop.

b. Balancing Foraging and Sense Making

The analyst is required to balance his time between gathering evidence and integrating it into a coherent story. During this process, the analyst usually has to deal with uncertainties, missing information and high complexity. It is also common to have to deal with contradictory evidence due to counter-intelligence, unreliable resources, misconceptions, etc.

The intelligence community has focused during the past decades on improving the intelligence collection capabilities in order to keep up with the information revolution and spread of communication technologies. As a consequence, intelligence analysis is far exceeded now by the collection capabilities, and is required to focus on detecting relatively few valuable information items and integrate them to support reasoning (Heuer, 2001).

If no classification of information as valuable or invaluable is made prior to the analysis, the amount of available information may be enormous and significantly larger than the capability of the analyst to digest. Hence, there is an acute need for a method to cull the valuable information to be analyzed and discard the rest. The challenge of reconciling the large collection capacity with the limited analysis capability has been addressed in recent studies, such as graph-based algorithms (Coffman et al., 2004), risk-based methodologies for scenario tracking (Horowitz & Haimes, 2003) and

Bayesian fusion of information (Paté-Cornell, 2002). These methodologies offer powerful tools to the analyst in performing the *sense-making loop* when coping with current analysis challenges. Greitzer (2005) tackles the challenge of assessing the impact of a tool or a methodology in the intelligence analysis context.

Information retrieval efforts (Singhal, 2001), which are widely applicable in civilian applications, constantly improve the ability to identify and retrieve the valuable information out of a large database. Nevertheless, if not bounded, the foraging efforts may take much of the analyst's time, resulting in less time available for sense-making, and a degraded overall performance.

In order to overcome this risk, the analyst may take advantage of the a-priori classification that the intelligence cycle offers at the processing phase. Early classification is implicitly employed when unprocessed or unusable information items are discarded at the processing phase, e.g., information that could not be translated, decrypted, etc. Nevertheless, early classification can also be performed in order to meet the analyst's requests regarding the desired items. Items that meet those requests are the ones to be processed and submitted to the shoebox before other items, thus their relative portion increases.

5. Dissemination

The final phase requires the distribution of the intelligence products to the consumers who initiated the corresponding IR. The disseminated intelligence may have different formats, such as reports, bulletins, assessments, studies etc., and may be distributed in different temporal manners: periodically, upon retrieval, upon request etc. Once the products are disseminated, the cycle goes back to planners and directors to re-prioritize IRs and issue new ones.

B. PROCESSING PHASE CLASSIFICATION

A key difference between the processing and analysis phases lies in the nature of the classification. While classification by an analyst is tentative since the item may be revisited during one of the following foraging iterations, classification by the processor is irreversible since items that are filtered out are discarded. Since analysis and processing are usually done by separate organizations (Defense Science Board, 2009), that difference only deepens due to immobility of resources between the phases and agenda conflicts. Figure 3 displays how items discarded in the processing phase, either before actual processing (top) or after (bottom), do not reach the repository that serves the analysis phase later on. The top figure displays a bottleneck at the processing phase and the bottom displays a bottleneck at the analysis phase.

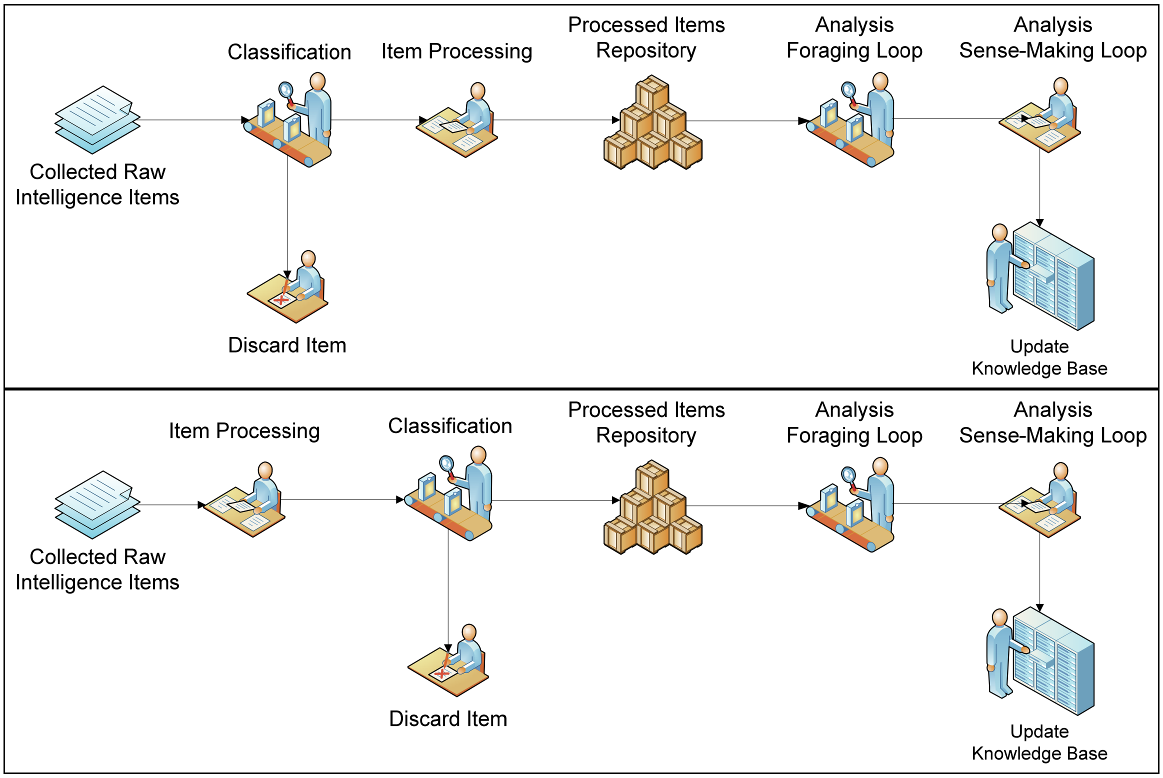
Classification
Item Processing
Processed Items Repository
Analysis Foraging Loop
Analysis Sense-Making Loop
Collected Raw Intelligence Items
Discard Item
Update Knowledge Base
Item Processing
Classification
Processed Items Repository
Analysis Foraging Loop
Analysis Sense-Making Loop
Collected Raw Intelligence Items
Discard Item
Update Knowledge Base

Figure 3. Pre-processing classification (top) due to a bottleneck at the processing phase, e.g., translation and Post-processing classification (bottom) due to a bottleneck at the analysis phase

If the analyst is only interested in a small portion of the collected information, processing all of it may result in tilting the balance between foraging and sense-making towards the first, resulting in a degraded overall performance. In particular, this is the case when each processed item is read by the analyst, or when the processed items repository has limited capacity, especially in SIGINT. In this case, the analyst may delegate the search and filter phase to the processing phase. Hence, even when the processing phase does not have an inherent bottleneck, e.g., when it is automated, a requirement for processor classification may arise in order to meet limitations at the analysis phases.

C. BINARY CLASSIFICATION PROCESS AND MEASURES

As pointed out in previous sections, the collected information goes through a binary classification process; an item is either classified as valuable to a certain IR and true in the sense that it conveys ground truth, or otherwise. Items that are declared as non-valuable or untrue are discarded, while the rest are processed and make their way to the foraging loop of the analysis phase. Since intelligence items are handled individually, the classification process becomes, in essence, a typical binary classification that assigns each item into two groups: valuable and non-valuable. The classification is done based on a test that comprises a set of evaluation questions and a decision mechanism that draws the line between a valuable item and a non-valuable one given the results of the test. For example, the decision on whether an IMINT product is valuable or not for a certain IR is naturally based on a test, which among other questions asks: is the image changed since the last processed image of the same area of interest; does it answer the IR; is it usable enough for analysis, etc.

1. Sensitivity and Specificity

In order to characterize the performance of a binary classifier, two statistical measures are widely used: Sensitivity and Specificity. If we refer to a valuable item as a "positive" and to a non-valuable item as a "negative," the sensitivity *p* is the probability of correctly classifying a positive, while the specificity *q* is the probability of correctly classifying a negative:

$$p = \Pr\left(positive\,id \mid positive\,item \right) \text{ and } q = \Pr\left(negative\,id \mid negative\,item \right)$$

The sensitivity and specificity of a classifier can be estimated by measuring empirically the performance of the classification on a set of items. Let *TP* be the number of true positives (i.e., positives correctly classified as positives), *TN* be the number of true negatives, *FP* be the number of false positives (i.e., negatives incorrectly classified as positives) and *FN* be the number of false negatives. In the experiment, the *TP*, *TN*, *FP* and *FN* are counted with respect to the experimenter's knowledge, and in our context, the analyst's.

Given these counts, the estimates for the sensitivity, *p*, and the specificity, *q*, are given by:

$$sensitivity = p = \frac{TP}{TP + FN} \text{ and } specificity = q = \frac{TN}{FP + TN} \quad (2.1)$$

The false positive rate (*FPR*) is the probability of incorrectly classifying a negative item, and the false negative rate (*FNR*) is the probability of incorrectly classifying a positive item. The specificity and sensitivity are the complement probabilities of the *FPR* and the *FNR*, respectively, That is:

$$FPR = 1 - q = \frac{FP}{FP + TN} \text{ and } FNR = 1 - p = \frac{FN}{TP + FN} \qquad (2.2)$$

A perfect classification is characterized by perfect sensitivity and specificity, i.e., $p = q = 1$. On the other hand, an arbitrary classification, with certain probability r of declaring an item as positive, is characterized by any sensitivity and specificity probabilities that satisfy $p = 1 - q = r$. For example, $p = q = 0.5$ characterize a classification system that uses a fair coin to decide whether an item is positive or negative, while $p = 1, q = 0$ implies no classification; all items are declared as positives.

Given a single evaluation question answered on an item, the classification system decides whether to declare the item as positive or negative using a decision mechanism. By narrowing this decision mechanism, e.g., changing a threshold value, the classification system can increase the sensitivity while decreasing the specificity and vice versa. The decision mechanism represents a policy, which is executed in the classification process. When a longer set of questions is answered on each item, more information is gathered; hence, a more informed decision can be made and the accuracy may improve. Therefore, there are essentially two degrees of freedom when planning a classification process: the test, namely the set of questions to be answered for each item, and the decision mechanism that depends on the possible outcomes of the test.

The accuracy, *ACC*, measures the overall rate of correct classification, estimated empirically as $ACC = \frac{TP + TN}{P + N}$, where $P + N$ is the total number of classified items. While *ACC* used to be a common measure of effectiveness for summarizing the performance of a binary classification system, it has also been criticized as a poor and misleading metric by Provost et al. (1998) mainly because it assumes equal costs for both misclassification types (*FP* and *FN*) and that the distribution of the positives and the negatives in the target environment is known. Instead of the single-number metric, they propose the use of Receiver Operating Characteristic curve.

2. Receiver Operating Characteristic (ROC) Curves

A widely used tool to explore and present the tradeoff between the sensitivity and specificity of a binary classification system is a graphical plot called *receiver operating characteristic* or, in short, *ROC curve* (Swets, 1988; Fawcett, 2006).

ROC curves characterize classification systems and allow comparison of systems as well as cost benefit analysis of the decision mechanism. These curves are common in machine learning, data mining, as well as in medicine and radar operation.

The x-axis of the ROC curve is the false-positive rate, $FPR = 1 - q$, and the y-axis is the sensitivity p. The *ROC space* is the (theoretically) feasible region of the ROC curve, namely the square region defined by the points *(0,0), (0,1), (1,1), (1,0)*. The point $(0,1)$ represents a *perfect classification* system with no errors. A point $(p_0, 1 - q_0)$ on the diagonal line $p = 1 - q$ represents a random guess with certain probability $r = p_0$ of declaring an item as positive.

A piecewise linear approximation of the ROC curve may be generated by a series of experiments with the classification system. For certain decision mechanisms, the system classifies a set of items and the probabilities are computed out of the classification results to create a single point on the ROC curve.

Because one can always reverse the cue, and assuming that the classifier is not biased due to deception or other measures, the random guess is the worst case scenario and therefore no $\left(1-q, p\right)$ values are possible under the diagonal $p=1-q$.

The area under the ROC curve (*AUC*) is a scalar measure of performance for the classification system. It allows a high-level comparison among different classification systems on the entire ROC space. The *AUC* can vary from 0.50 (worst-case scenario) to 1.0 (perfect classification) when assuming that no classification system can go below the diagonal (Marzban, 2004).

The adjustment of the ROC curve of an intrusion detection system given a limited investigation capacity is explored in the field of computer networks security by Yue and Cakanyildirim (2010). They use a decision tree approach in which the cost of each investigation is weighted against the damage by such an intrusion. This type of cost-based analysis is hard to apply under the context of intelligence since the benefit of a single item is measurable only with respect to the final product of the cycle, rather than its value as an independent item. Therefore, the notion of damage per unidentified item is artificial.

3. Precision and Recall

A closely related area of research is Information Retrieval, which is the science of searching for documents corresponding to a certain query in a set of general documents, usually held in a database or another repository. Information Retrieval uses similar performance measures for the system, namely *precision* and *recall*. *Precision* is the fraction of valuable documents that are retrieved among the total number of documents the same search retrieved. *Recall* is the fraction of valuable documents retrieved out of the total number of valuable documents in the search set. For an empirical trial, the *precision* and *recall* of a classification system are estimated using the same parameters as before, and their relationship to the sensitivity, *p*, and the specificity, *q*, is given by:

$$precision = \frac{TP}{TP+FP} = \frac{p \cdot P}{p \cdot P + \left(1-q\right) \cdot N} \text{ and } recall = \frac{TP}{TP+FN} = p \qquad (2.3)$$

where *TP, FP* are the number of true positives and false positives, respectively, *FN* is the number of false negatives, and *P*,*N* are the total number of positives and negatives in the trial set.

While both measures were originally developed to characterize set retrieval, current research deals with ranked retrieval models that ranks results by estimated likelihood of relevance to the given query. Among the measures that take into account the ranked order of the results are mean average precision, MAP, and normalized discounted cumulative gain, NDCG (Järvelin & Kekäläinen, 2002).

For information foraging problems that aim at very rare but valuable information, recall is still used as the primary performance measure. In the extreme case of information foraging, the problem becomes an *information availability* problem in which the information seeker is uncertain regarding the very existence of the information that was searched for. In these problems, which often include intelligence analysis situations, the importance of recall as a measure of effectiveness becomes critical (Pirolli, 2009).

Similar to the ROC curves used in the sensitivity-specificity terminology, *Precision-Recall* (PR) curves are used in fields such as machine-learning and data mining in order to illustrate graphically the performance of a classification system. One advantage of the ROC curve over the PR-curve is its independence of the *P* and *N* values that are required to compute the precision. Nevertheless, the PR curve is equivalent to the ROC curve, in the sense that a domination relationship between curves in the ROC space exists if and only if it exists in the PR space (Davis & Goadrich, 2006).

D. INTELLIGENCE OPERATIONS RESEARCH

Intelligence Operations Research (OR) refers to the implementation of OR techniques to benefit and improve the intelligence process, by modeling and solving specific problems of the intelligence community. A broad overview of the subject is provided by Kaplan (2010b).

The “Advisory Group on Defense Intelligence,” which is a committee of the Defense Science Board to examine and advise on matters related to defense intelligence in the DoD, examined the use of OR in Intelligence, Surveillance and Reconnaissance, ISR (Defense Science Board, 2009) and concluded that “Operations Research is applied inconsistently throughout the Defense and ISR communities. These communities do not possess standard OR processes and practices, a consistent organizational model, or a consistent commitment to the use of OR” (page 37).

In order to establish OR as a beneficial methodology for intelligence, two test cases were suggested: (1) Balancing investments in the intelligence cycle (2) Investment decision making in biometrics technologies. The significant challenge that the task force pointed out is how to formulate the objective function of the intelligence product and its production phases. Another issue was the diversity of organizations in charge of the various phases of the intelligence production process, making cross-phase resource allocation more difficult to implement.

Few operations research models were suggested for the intelligence work itself. Steele (1989) models the time as until a secret is disclosed, when shared among members of a group.

Skroch (2005) uses an interdiction model to optimize interdiction resources in order to hinder a nuclear weapons project. Another proposed model (Pinker et al., 2009) is a mixed integer linear programming model in which the proliferator minimizes the time from detection to completion rather than simply minimizing time to completion.

Kaplan (2010a) describes models for infiltration and interdiction of terror plots by HUMINT agents using “terror queues,” Markovian queues in which terror plots are customers and the HUMINT agents are the servers interdicting the served terror plots. The reneging terror plots are those that are executed before being interdicted.

III. THE MODEL

A. CHAPTER OVERVIEW

This chapter presents the model of the intelligence processing-analysis system described in Chapter II and discusses its assumptions. The proposed model is a tandem queue where each station corresponds to a phase in the system: processing and analysis. The basic queuing model is implemented in an optimization model that determines operational parameters.

As discussed in Chapter II, many research efforts have been focused on both the processing and analysis phases of the intelligence cycle, proposing qualitative and quantitative methods to overcome the challenges associated with the operation of each phase. Nevertheless, to the best of our knowledge, none of these studies focused on quantifying operational parameters and allocating resources between the two key phases in the cycle: processing and analysis. Even if efficient and effective practices are implemented in each phase separately, it may well be that as a combined system the operation is not optimal. Overcoming this possible sub-optimality is the main motivation for the model described in this chapter. The main goal of our model is to determine the optimal values of the tactical parameters of the classification and compare the improvement of the optimal setting under different scenarios. To keep the model general we consider a highly aggregated form of each phase, focusing on easy to measure characteristics of the phase, such as service rates and quality of performance, rather than on its detailed modus operandi.

B. OPERATIONAL SETTING

Before formulating the model, we describe the operational setting and pose some assumptions. The basic scenario considers a single *Intelligence Requirement* (IR) and a given *collection plan,* which determine the characteristics of the *intelligence items* for processing. For the sake of brevity, we simply refer to those as *items.* The *items* that are going through the processing phase consist of two types with respect to the specified IR: valuable items, called henceforth positives (*P*), and worthless items, called henceforth negatives (*N*). Positives are items that are both relevant to the IR and convey additional ground truth according to the analyst, and negatives are items that are either irrelevant to the IR, repetitive, or contain incorrect information, again, according to the analyst.

As discussed in Chapter II, the process we model is described as follows: an incoming item is first classified as positive or negative based on the information gathered on that item by the processor.

If the item is classified as a positive then it is passed on to the analysis phase; otherwise, the item is discarded and cannot be revisited later on. The analyst then inspects each submitted item and establishes its significance and implication on current knowledge base, e.g., by updating the IR or disseminating the newly retrieved information to relevant decision makers.

The classification at the first phase is subject to false-negative and false-positive errors (*FN* and *FP*). While false-positive items increase the number of items that are passed on to analysis, and thus increase the load for the analysts, they will eventually be detected as such by the analysts and therefore will cause no additional harm. On the other hand, the false-negative items comprise valuable information that is lost, at least until it may reappear in another positive item. Figure 4 describes the processing-analysis system.

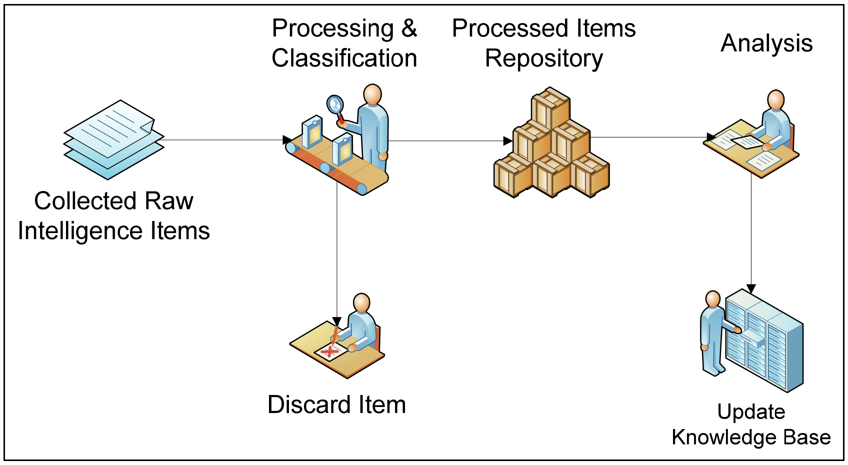

Figure 4. The processing-analysis system

C. THE TANDEM QUEUE MODEL

Recall that we assume that information arrives in a discrete form of *items*. The stochastic counting process of items arriving for processing is the sum of two independent homogeneous Poisson processes: (1) positives $\{P(t);t \geq 0\}$ with an arrival rate λ_P and (2) negatives $\{N(t);t \geq 0\}$ with an arrival rate λ_N. Given an item *i*, we denote $i \in P$ if it belongs to $P(t)$, and $i \in N$ otherwise.

Let *j=1,2* be the index of the processing station and the analysis station, respectively, and let $\lambda_j, \mu_j \in (0,\infty)$, *j=1,2*, be the arrival and service rates at station *j*.

Assuming that the inter-arrival and service times are exponentially distributed and independent of everything, the resulting model is a tandem M/M/1 queue. Given the arrival process rates λ_P, λ_N, the *sensitivity p* and *specificity q* of the classification process, the arrival rate of items for analysis is:

$$\lambda_2 = p\lambda_P + (1-q)\lambda_N \qquad (3.1)$$

The tandem queue is displayed in Figure 5. Note that while the arrival processes are displayed separately in the figure for clarity purposes, the processing station cannot separate between positives and negatives before service completion, and sees both as a single arrival processes. In addition, the independence assumption implies that previous classifications have no effect on the true classification of incoming items, in the sense that there is no cumulative learning from the intelligence products.

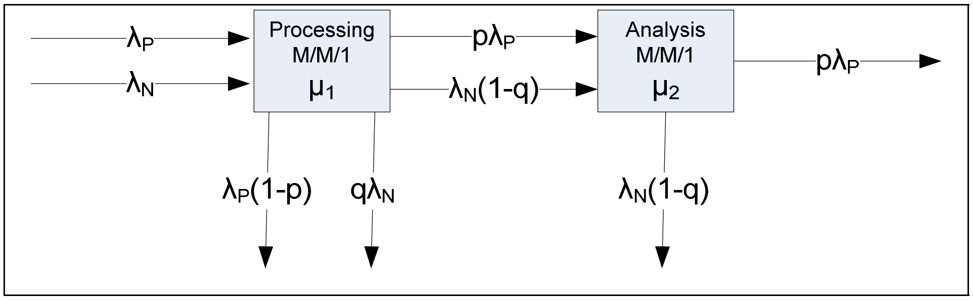

Figure 5. The tandem M/M/1 model for the described process

The tandem queue is stable if $\lambda_j < \mu_j$ for both stations *j=1,2*, in which case W_j, the long-run expected delay of an item in station *j*, is finite. The long-run expected delay of an item in the processing-analysis system, *W*, is given by:

$$W = W_1 + W_2 = \frac{1}{\mu_1 - \lambda_1} + \frac{1}{\mu_2 - \lambda_2} = \frac{1}{\mu_1 - \lambda_1} + \frac{1}{\mu_2 - p\lambda_P - (1-q)\lambda_N} \tag{3.2}$$

While the average service rate at the analysis station, μ_2, only affects W_2, the average service rate of the processing phase, μ_1, affects both W_1 and the classification quality, manifested in *p* and *q*, which affects λ_2 and therefore also W_2. Introducing a relationship between the classification quality and the rate in which it is performed allows us to capture tradeoffs between quality and quantity in the classification process at station 1. Decreasing μ_1, that is slowing down the classification, increases W_1 and allows performance of a longer test; we assume the quality of the classification increases as well. When μ_1 increases, the opposite holds. In reality, increasing μ_1 may be manifested, for example, by reducing the number of questions asked on each item in the processing phase, or answering these questions faster. We assume that a longer mean service time $\frac{1}{\mu_1}$ is associated with a longer sequence of questions that the classifier is able to answer on each item; this will result in a higher classification accuracy. In the next section, we discuss the modeling of the relationship between μ_1 and the quality of the classification as manifested in its sensitivity and specificity.

D. THE CLASSIFICATION PROCESS

As mentioned above, given an item, the classification process can be abstractly modeled as a sequence of questions used to gather information about that item. Once those questions are answered, a decision mechanism decides, based on the collected information embodied in the answers, whether the item should be declared as positive or negative. We define a *test* as a set of questions asked about each classified item. For example, for a test with a single question, which results in a scalar describing the item, the decision mechanism may take the form of a threshold value that classifies an item as positives if the scalar value is higher than the threshold and negative otherwise.

1. Modeling the Classification

To characterize the classification process implemented in the processing phase, we define the *binary classification setting* as a vector of two elements: (1) the implemented *test* and (2) the decision mechanism used to declare an item as positive based on the test results. The method by which each one of the two elements is controlled in an operational environment is discussed in the subsequent sections.

Each possible *test* can be used to plot empirically the *ROC curve* that describes the relations between the sensitivity $p = \Pr\{TP \mid P\}$ of the test and its specificity $q = \Pr\{TN \mid N\}$ as one changes the decision mechanism, as discussed in Chapter II. Recall from Chapter II that the ROC curve is the sensitivity as a function of the false positive rate: $p = f(1-q)$.

For each *test*, the resulting ROC curve of the classification starts at *(0,0)*, which represents the trivial threshold where all items are declared as negatives, and ends at *(1,1)*, which represents the trivial threshold where all items are declared as positives. We assume that the ROC curve is given in a general parametric form:

$$p = f_{\varepsilon}(1-q,\varepsilon) \text{ for some parameter } \varepsilon \in [\varepsilon_{\min}, \varepsilon_{\max}] \qquad (3.3)$$

Substituting (3.3) in the expression for λ_2 given by (3.1) yields:

$$\lambda_2 = f_{\varepsilon}(1-q,\varepsilon)\lambda_P + (1-q)\lambda_N \qquad (3.4)$$

The parameter ε is a measure of the quality of the classification. Without the loss of generality, when ε assumes the value of its upper bound, $\varepsilon_{\max}$, it implies the worst-case scenario of random classification. This scenario occurs when, regardless of its characteristics, an item is classified as *P* with a certain probability *r* and as *N* with probability 1-*r*. In that case the corresponding ROC curve is the diagonal $p = 1-q = r$. On the other hand, when ε reaches its lowest possible value, i.e., $\varepsilon = \varepsilon_{\min}$, the best possible classification is achieved, that is $p = f_\varepsilon(1-q,\varepsilon_{\min}) = 1$ for all q, and in particular for $q = 1$, perfect classification is achieved. We assume that when the test is expanded, i.e., more questions are answered on each item, ε decreases. Figure 6 illustrates the above formulation in the ROC space for $f_\varepsilon(1-q,\varepsilon) = (1-q)^\varepsilon$ where $\varepsilon \in [0,1]$.

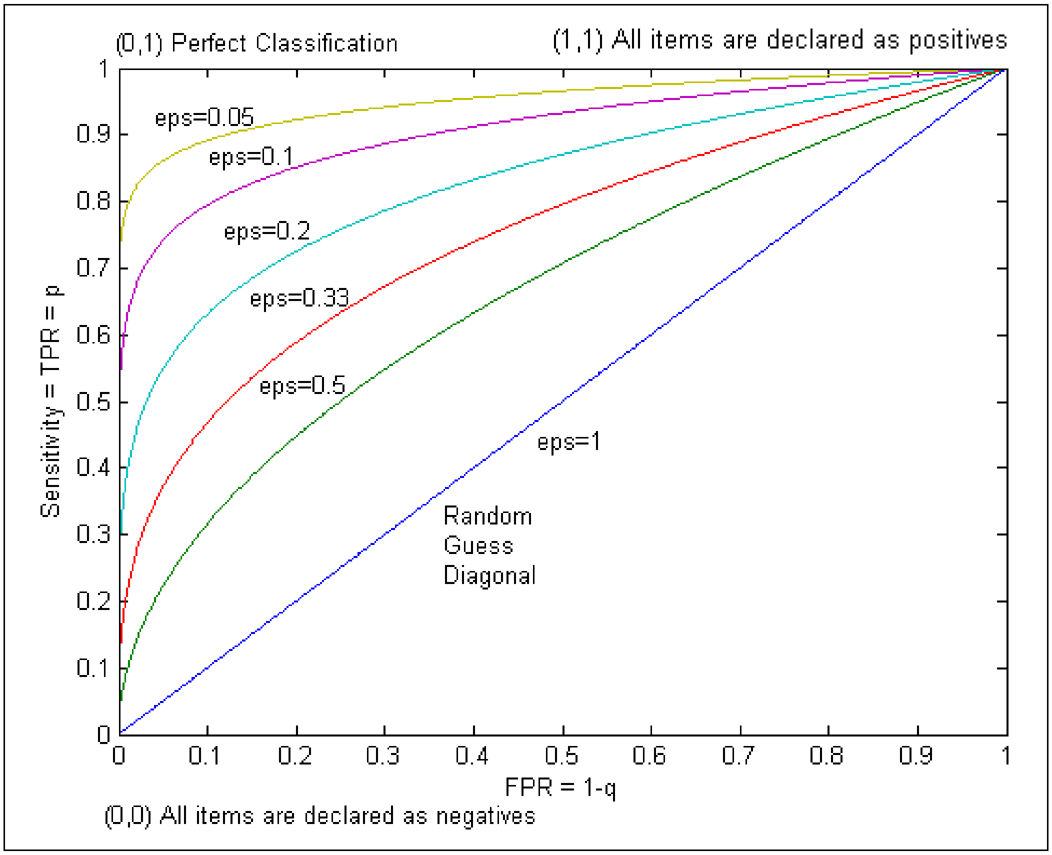

Figure 6. ROC Space diagram for $f_\varepsilon(1-q,\varepsilon) = (1-q)^\varepsilon$ and different values of ε.

2. Controlling the Classification Setting

Recall that the classification process is defined by two elements: (1) the *test* performed on each item before it is classified, and (2) the decision mechanism used to declare an item as positive. In this section, we discuss the method by which each element can be adjusted.

Assume that the *test* performed on each item before classification is a subset of some global set of evaluation questions, each associated with the average time it takes to answer the questions and its corresponding ROC curve, which represents different decision mechanisms based on the information accumulated during the test. For example, we mentioned three questions with respect to the example on IMINT products described in Chapter II: (1) how suitable it is for analysis (e.g., quality of the image and its resolution), (2) how strongly is the image related to the IR, and (3) how significant is the change in the image since the last time the area was visited. Based on these questions one can theoretically create 2^3 tests; e.g., only question 1; or only questions 2 and 3, etc. However, in reality, four tests might be sufficient since we can order the questions by some measure of effectiveness from 1 to 3 and then test *i* contains questions with rankings 0 through *i*. Test 0 contains no questions, and test 3 contains all of them. For each of those tests the service rate and the ROC curve can be computed empirically, by utilizing several decision mechanisms. In our example, test 2, which contains questions (1) and (2), may have a two-dimensional threshold mechanism, in which both the quality and the relation to the IR should exceed certain reasonable levels.

Let the *empty test* be the test in which no question is included and let the *complete test* be the test in which all of the available questions in the global set are included. When performing the *empty test*, no information is collected on the item; therefore, the best classification is given by a random guess ROC curve. Naturally, the more questions one asks on each item, the better is the classification of items to either *P* or *N*, since more information is gathered to support the decision, and we assume no deliberate disinformation. Thus, the value of ε in the ROC curve does not increase when more questions are answered. When additional questions add no more information, the value of ε stays fixed.

When the service rate at the processing station, μ_1, increases due to reduction of the test length, that is, classifiers are instructed to process each item faster by addressing fewer questions, ε gets larger, and the resulting ROC curve “shrinks” south-east, towards the random guess diagonal. Thus $\varepsilon\left(\mu_1\right)$ is assumed monotone increasing in μ_1. In addition, since it is reasonable to assume that relatively ineffective questions, i.e., questions that contribute less to the classification of the ROC curve with respect to the time they consume, may be removed first; $\varepsilon\left(\mu_1\right)$ is also assumed to be concave.

The measurement of $\varepsilon\left(\mu_1\right)$ requires a thorough discussion. For a predefined test, that is, a sequence of questions fully addressed on each item, an experiment should include the measurement of both the mean time it takes to perform the test $\left(\mu_1^{-1}\right)$ and the corresponding ROC curve. The measurement of the mean time should refer to operationally achieved mean time, e.g., by excluding learning effects. The ROC curve can be empirically drawn using different decision mechanisms applied to the same information collected during the performance of the test. Given the empirical plot of the ROC curve, the best fitted value of ε should be computed to approximate the retrieved curve as the functional form $f_\varepsilon\left(1-q,\varepsilon\right)$. Then, given these results for multiple tests, the functional form $\varepsilon\left(\mu_1\right)$ may be estimated.

Let $\underline{\mu}_1 \geq 0$ be the highest value of μ_1 for which the *complete test* can be performed, meaning all possible questions in the global set can be addressed. In that case the system reaches the optimal classification capability – the minimal value of ε. Let $\underline{\varepsilon}$, $\varepsilon_{\min} \leq \underline{\varepsilon} \leq \varepsilon_{\max}$, be the *classification capability limit*, that is $\varepsilon\left(\underline{\mu}_1\right) = \underline{\varepsilon}$. On the other hand, when μ_1 becomes high enough such that no test can be performed, the classification reaches its worst performance – a random guess. Let $\bar{\mu}_1 > \underline{\mu}_1$ be the smallest value of μ_1 such that $\varepsilon\left(\bar{\mu}_1\right) = \varepsilon_{\max}$.

For example, a linear relationship that satisfies the above two conditions is:

$$\varepsilon\left(\mu_1\right) = \frac{\mu_1 - \underline{\mu}_1}{\bar{\mu}_1 - \underline{\mu}_1}\left(\varepsilon_{\max} - \underline{\varepsilon}\right) + \underline{\varepsilon} \qquad (3.5)$$

The bounds $\bar{\mu}_1$ and $\underline{\mu}_1$ are estimated by the longest average time no other test can be performed except for the *empty test* and by the shortest average time to perform the *complete test*, respectively. The expected service time is $\frac{1}{\mu_1}$. The upper bound $\bar{\mu}_1$ takes into account the overhead time per item at the classification station. For example, some internal administrative procedures may be required at the processing station regardless of the quality of the classification.

The value of $\underline{\varepsilon}$ is estimated from the *ROC curve* of the classification when the *complete test* is performed. Note that the region $\left[\varepsilon_{\min}, \varepsilon_{\max}\right]$ is the mathematical region for which the ROC curve is feasible, while the region $\left[\underline{\varepsilon}, \varepsilon_{\max}\right]$ is the achievable region by the classification.

By substitution of $\varepsilon\left(\mu_1\right)$ in (3.3) and (3.4) respectively, we have:

$$p = f\left(1-q, \mu_1\right) \text{ for } \mu_1 \in \left[\underline{\mu}_1, \bar{\mu}_1\right] \qquad (3.6)$$

$$\lambda_2 = f\left(1-q, \mu_1\right)\lambda_P + \left(1-q\right)\lambda_N \qquad (3.7)$$

This formulation uses two sets of variables. The first set comprises the strategic parameters of the classification phase: the range of classification work intensities $\left[\underline{\mu}_1, \bar{\mu}_1\right]$, and the *classification capability limit* $\underline{\varepsilon}$. These characteristics can be controlled: the bounds $\bar{\mu}_1$ and $\underline{\mu}_1$ can be increased by training the classifiers and performing the test more efficiently, and the *classification capability limit* $\underline{\varepsilon}$ can be decreased by generating questions with a better ability to classify the items.

The second set consists of the tactical variables μ_1, p and q, that decide the modus operandi in which the processing phase classifies items. These variables give two degrees of freedom for policy makers; first, the choice of μ_1 determines the shape of the ROC curve by determining the value of $\varepsilon\left(\mu_1\right)$. Then the choice of *p* determines the value of q on the ROC curve, and vice versa. Without an analysis station bottleneck, one would presumably set $p=1-q=1$ for every value of μ_1; however in the presence of limited analysis resources and delay constraints, this setting becomes infeasible.

In this thesis, we focus on the optimization of the classification by setting the value of the tactical decision variables, where the impact of the input variables is further explored in the analysis presented in Chapter IV.

The value of *p* can be set by relaxing or restricting the decision mechanism for declaring an item as positive. To illustrate the adjustment of *p* and *q*, suppose the analysts are looking for items associated with a certain subject. Consider, for example, a one-dimensional test that, given a dictionary of terms, counts the number of term occurrences in each item. For that specific test example, the value of ε depends on the quality of the dictionary of terms. The better the dictionary, i.e., the better it differentiates between positives and negatives, the value of ε is lower. The decision mechanism is a simple threshold on the count. If we require zero occurrences of terms to declare the item positive, the classification achieves $p=1-q=1$, which represents a random guess with $\Pr(P)=1$. As we set the threshold higher, the number of items that meet the threshold decreases, including both positives and negatives. Therefore, the sensitivity *p* decreases since the false negative rate increases while the specificity *q* increases since the false positive rate decreases. At a certain point we may require more occurrences of terms than there exist, resulting in $p=1-q=0$, which is a random guess with $\Pr(P)=0$.

Suppose the dictionary contains *n* terms. If we generalize the test and count the number of occurrences for each one of the *n* terms, the decision mechanism becomes an *n*-dimensional function of the counts.

3. The Alternative of No Classification

In the case where classification is not implemented at all, the analysis station inspects the material in its raw form. We assume the worst-case scenario where the inspected items are randomly chosen from the flow of arriving items with a certain probability *r*, according to the analysis station service rate. Thus the system is now a single M/M/1 queue where $W = \frac{1}{\mu_2 - r\lambda_1}$, in which λ_1 is the arrival rate of all items to the processing server, both positives and negatives. Therefore, the sensitivity satisfies $p_{\min} = r$ and can be expressed as:

$$p_{\min} = \frac{\mu_2 - \frac{1}{W}}{\lambda_1} \qquad (3.8)$$

Expression (3.8) represents the lower bound on the sensitivity that may be achieved by the system, thus the notation $p_{\min}$.

E. OBJECTIVES

As discussed in Chapter II, intelligence products are commonly measured by quality, quantity, timeliness, and information needs satisfaction. Since we focus in this study only on the part of the intelligence process in which the final product is not yet tangible, the quality and information needs satisfactions are hard to estimate.

Therefore, our measure of effectiveness is the sensitivity *p* of the system (also known as the recall), given an upper bound on the acceptable total expected delay *W*, which affects the timeliness of the intelligence product.

This concludes the formulation of the framework for the basic model. The following sections present the optimization model with respect to the stated objective, which we call the "Classification Optimization Model." The classification optimization model deals with the optimal setting of the classification at the processing phase, given that the arrival rates and the analysis service rate are fixed. In other words, the model deals with these questions: what should the service rate μ_1 and the point $\left(1-q, p\right)$ on the ROC curve corresponding to $\varepsilon\left(\mu_1\right)$ be? In addition, the model allows sensitivity analysis for the values of the strategic variables.

1. Cost-Effectiveness of the System

When it comes to intelligence processing it is reasonable to choose the sensitivity (*p*) as the objective of the system. However, given the optimal performance of the system, one would like to measure its cost-effectiveness as well for decision making purposes. As was discussed in Chapter II, it is hard to quantify the value of an independent intelligence item in the wider context of the intelligence product. On the other hand, using the framework presented in this chapter, we can formulate expectancy based measures for the total cost of the coupled system.

Let the cost of an analyst per unit time be one budget unit, and let *b* be the classifier cost per unit time. Since the best classifier is the analyst, the range of the classifier cost is given by $0 \le b \le 1$.

The expected cost of an item declared as negative at the classification phase is $\frac{b}{\mu_1}$ while the expected cost of an item declared as positive costs $\frac{b}{\mu_1}+\frac{1}{\mu_2}$ since it goes through both stations. Therefore, the expected cost of the coupled system is given by $b\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2}$. On the other hand, when no classification is performed, the expected cost is given by $\frac{p_{\min}\lambda_1}{\mu_2}$, where $p_{\min}$ is given by (3.8), since each one of the $p_{\min}\lambda_1$ processed items requires an expected service time of $\frac{1}{\mu_2}$.

In order to incorporate the performance of the system into the cost measure, we look at the expected cost per correctly identified positive. Let *B* be the expected total cost of the system per positive, meaning the cost of both the classification and the analysis over a single period divided by the number of correctly identified positives processed during that period. For a coupled system the number of correctly identified positives is given by $p\lambda_P$ (see 3.1); thus we have $B = \frac{b\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2}}{p\lambda_P}$ and for a system without classification we have $B = \frac{\frac{p_{\min}\lambda_1}{\mu_2}}{p_{\min}\lambda_P} = \frac{\lambda_1}{\mu_2\lambda_P}$.

We use the cost per correctly identified positive, *B*, to compare the cost-effectiveness of classification systems in different scenarios, when the classification setting optimizes the sensitivity.

F. CLASSIFICATION OPTIMIZATION MODEL

In this section, we formulate the model used to determine the optimal setting of the classification at the processing phase, as manifested by the variables p, q and μ_1, given the strategic parameters and the input parameters $\lambda_P, \lambda_N, \mu_2$, and the functional relationship $\varepsilon(\mu_1)$.

In the model we maximize the sensitivity *p* subject to the constraints explained previously, namely: (1) stability constraints on both stations in order to assure that each station is able to serve the flow of incoming items, (2) requirements that the sensitivity and specificity pair lies on the ROC curve, (3) the service rate at station one, μ_1, is restricted to be between the two bounds $\underline{\mu}_1$ and $\overline{\mu}_1$, and (4) an upper bound on the acceptable total expected delay $\overline{W}$. Formal definition of the problem is given in (3.9):

$$\max_{p,q,\mu_1,W} p \qquad (3.9)$$

Subject to:

- $\mu_j \geq \lambda_j \; j = 1, 2$ [Stability constraints]
- $0 \leq p, q \leq 1$ [ROC Space constraints]
- $\mu_1 \in \left[\underline{\mu}_1, \bar{\mu}_1\right]$ [Classification rate range]
- $W \leq \bar{W}$ [Delay upper bound]

In the subsequent section, we show that the model can be reduced to a two-dimensional optimization problem in μ_1 and q by expressing the objective as a function $p\left(\mu_1, q\right)$ following (3.6), substituting W following (3.2), and substituting $\lambda_2 = f\left(1-q, \mu_1\right)\lambda_P + \left(1-q\right)\lambda_N$ following (3.7). Therefore the problem is given:

$$\max_{q, \mu_1} f\left(1-q, \mu_1\right)$$

Subject to:

- $\mu_1 \geq \lambda_1$ [Stability constraint on station 1]
- $\mu_2 \geq f\left(1-q, \mu_1\right)\lambda_P + \left(1-q\right)\lambda_N$ [Stability constraint on station 2] (3.10)
- $0 \leq q \leq 1$ [ROC Space constraints]
- $\mu_1 \in \left[\underline{\mu}_1, \bar{\mu}_1\right]$ [Classification rate range]
- $\dfrac{1}{\mu_1 - \lambda_1} + \dfrac{1}{\mu_2 - f\left(1-q, \mu_1\right)\lambda_P - \left(1-q\right)\lambda_N} \leq \bar{W}$ [Delay upper bound]

IV. ANALYSIS AND RESULTS

A. CLASSIFICATION OPTIMIZATION MODEL ANALYSIS

Recall the classification optimization model of (3.10):

$$\max_{q,\mu_1} f\left(1-q,\mu_1\right)$$

Subject to:

- $\mu_1 \geq \lambda_1$ [Stability constraint on station 1]
- $\mu_2 \geq f\left(1-q,\mu_1\right)\lambda_P - (1-q)\lambda_N$ [Stability constraint on station 2]
- $0 \leq q \leq 1$ [ROC Space constraints]
- $\underline{\mu_1} \leq \mu_1 \leq \overline{\mu_1}$ [Classification rate range]
- $\dfrac{1}{\mu_1 - \lambda_1} + \dfrac{1}{\mu_2 - f\left(1-q,\mu_1\right)\lambda_P - (1-q)\lambda_N} \leq \overline{W}$ [Delay upper bound]

Throughout we assume that $f\left(1-q,\mu_1\right)$ is continuously differentiable and, without loss of generality, that $\overline{\mu}_1 \geq \lambda_1$, because otherwise no classification rate is feasible. As discussed in Chapter III, $f\left(1-q,\mu_1\right)$ is assumed non-decreasing in $1-q$ and non-increasing in μ_1, with boundary conditions $f\left(1,\mu_1\right)=1$, $f\left(0,\mu_1\right)=0$, and $f\left(1-q,\overline{\mu}_1\right)=1-q$. Observe that the expected delay is non-increasing in μ_1 and non-decreasing in $1-q$, and that at optimality we must have $\mu_1 > \lambda_1$ and $\mu_2 > f\left(1-q,\mu_1\right)\lambda_P - (1-q)\lambda_N$, for otherwise the delay constraint is violated for any $\overline{W}$ finite. Therefore both of these constraints are slack at optimality.

To get started, consider the random classification solution, where $\mu_1 = \overline{\mu}_1$, and $p = 1-q$. For $\mu_2 \geq (1-q)\lambda_1$, setting $p = 1-q$ is feasible whenever

$(\bar{\mu}_1-\lambda_1)^{-1}+(\mu_2-(1-q)\lambda_1)^{-1}\le\bar{W}$, meaning that random classification results in $p=\min\{\lambda_1^{-1}[\mu_2-(\bar{W}-(\bar{\mu}_1-\lambda_1)^{-1})^{-1}],1\}$ if $(\bar{\mu}_1-\lambda_1)^{-1}+\mu_2^{-1}\le\bar{W}$. The latter shows that $p\to\lambda_1^{-1}(\mu_2-\bar{W}^{-1})<1$ as $\bar{\mu}_1\to\infty$. Finally, random classification is unfeasible for $\bar{W}<(\bar{\mu}_1-\lambda_1)^{-1}+\mu_2^{-1}$ and optimal ($p=1$) whenever $\bar{W}\ge(\bar{\mu}_1-\lambda_1)^{-1}+(\mu_2-\lambda_1)^{-1}$ and $\mu_2\ge\lambda_1$.

For general classification schemes represented by $f\left(1-q,\mu_1\right)$, we also have that $\bar{W}<\mu_2^{-1}+(\overline{\mu_1}-\lambda_1)^{-1}$results in an unfeasible classification optimization problem, with $\mu_1=\bar{\mu}_1, q=1$ the only feasible solution for $\bar{W}=(\overline{\mu_1}-\lambda_1)^{-1}+\mu_2^{-1}$.

Let us now consider the remaining range of delay constraints, $(\overline{\mu_1}-\lambda_1)^{-1}+\mu_2^{-1}<\bar{W}$, where the problem is feasible. The standard theory suggests that the Karush–Kuhn–Tucker (KKT) conditions are necessary with, as discussed earlier, the delay constraint tight at optimality. It is possible, however, that the classification range constraints are tight as well. We denote $\nabla_x f=\dfrac{\partial f}{\partial x}$. Hence, the necessary conditions yield three candidate solutions:

- A stationary point with $\underline{\mu}_1<\mu_1<\bar{\mu}_1$ and KKT multiplier,

$$\nabla_q f\left(1-q,\mu_1\right)=\alpha\nabla_q\left(\frac{1}{\mu_1-\lambda_1}+\frac{1}{\mu_2-f\left(1-q,\mu_1\right)\lambda_P-(1-q)\lambda_N}\right) \tag{4.2a}$$

$$\nabla_{\mu_1} f\left(1-q,\mu_1\right)=\alpha\nabla_{\mu_1}\left(\frac{1}{\mu_1-\lambda_1}+\frac{1}{\mu_2-f\left(1-q,\mu_1\right)\lambda_P-(1-q)\lambda_N}\right) \tag{4.2b}$$

$$\frac{1}{\mu_1-\lambda_1}+\frac{1}{\mu_2-f\left(1-q,\mu_1\right)\lambda_P-(1-q)\lambda_N}=\bar{W}\,. \tag{4.2c}$$

- The random classification solution discussed earlier, $p=\min\{\lambda_1^{-1}[\mu_2-(\bar{W}-(\bar{\mu}_1-\lambda_1)^{-1})^{-1}],1\}$, where $\mu_1=\bar{\mu}_1$.

- The classification range constraint is tight at the lower bound, $\mu_1 = \underline{\mu}_1$, with $p = f(1-\underline{q}, \underline{\mu}_1)$, where $0 \le \underline{q} \le 1$ is as small as possible subject to making the delay constraint tight. We should point out that this solution cannot be optimal for $\underline{\mu} < \lambda_1 + (\bar{W} - \mu_2^{-1})^{-1}$ because in that range the delay constraint is not tight for any feasible q.

From these observations, we gather that random classification tends to perform very well when $\bar{W}$ and μ_2 are relatively large. Unfortunately, it is difficult to obtain other qualitative insights without making stronger assumptions about $f\left(1-q,\mu_1\right)$. This is precisely the approach we take in the next section, where we implement the model for a particular function $f\left(1-q,\mu_1\right)$.

B. CLASSIFICATION OPTIMIZATION MODEL: NUMERICAL EXAMPLE

This section presents numerical results from implementing the classification optimization model on various scenarios. We study the effect of the main parameters on the performance of the system, highlighting key insights that can be drawn both from the change in the optimal value and the arguments that achieve this optimal value.

The implementation of the model is done using the General Algebraic Modeling System (GAMS), which provides a high-level language for the purpose of implementing mathematical programming models (Brooke et al., 1998).

1. The Model

Recall from Chapter III that the ROC curve considered here is determined by two functional relations: $p = f\left(1-q,\varepsilon\right)$ for some parameters $\varepsilon \in \left[\varepsilon_{\min}, \varepsilon_{\max}\right]$ (see 3.3), and the relation between the parameter ε and the classification rate μ_1: $\varepsilon\left(\mu_1\right) = \varepsilon$.

In this implementation of the model we assume that $f\left(1-q,\varepsilon\right) = \left(1-q\right)^{\varepsilon}$ where $\varepsilon \in \left[0,1\right]$ (see Figure 6 for illustration) and that $\varepsilon\left(\mu_1\right) = \frac{\mu_1 - \underline{\mu}_1}{\bar{\mu}_1 - \underline{\mu}_1}\left(1-\underline{\varepsilon}\right) + \underline{\varepsilon}$ (see 3.5 for derivation), where $\underline{\varepsilon}$ is the *classification capability limit*, the lowest achievable value of the parameter ε. Let $a \equiv \frac{1-\underline{\varepsilon}}{\overline{\mu}_1 - \underline{\mu}_1}, b \equiv \frac{\underline{\varepsilon}\overline{\mu}_1 - \underline{\mu}_1}{\overline{\mu}_1 - \underline{\mu}_1}$, then $\varepsilon(\mu_1)$ can be written as $\varepsilon(\mu_1) = a\mu_1 + b$, and $f(1-q,\varepsilon) = (1-q)^{\varepsilon} = (1-q)^{a\mu_1+b}$.

2. Input Parameters

Recall that the input parameters to the models are the arrival rates, λ_P and λ_N of the positives and negatives, respectively, and the characteristics of the system, which include the analysis service rate μ_2, the classification rate range $\left[\underline{\mu}_1, \overline{\mu}_1\right]$, the acceptable delay $\overline{W}$, and the classification capability limit, $\underline{\varepsilon}$.

We consider three scenarios with respect to the arrival rates, all of which satisfy $\lambda_1 = \lambda_P + \lambda_N = 100$. A *low value source* has $\lambda_P = 1$, a *medium value source* has $\lambda_P = 10$ and a *high value source* has $\lambda_P = 20$. The analysis service rate is assumed to be at a constant level of $\mu_2 = 20$ for all three scenarios and the range of the classification rate is $[0, 500]$. The lower bound of zero represents unlimited number of answers that can be addressed with respect to each item, and the upper bound of 500 represents the average number of items that can be passed on to the analysis phase when no classification is made at all. We let the *classification capability limit* assume its best value - $\underline{\varepsilon} = 0$, thus $\varepsilon(\mu_1) = a\mu_1 + b = 0.002\mu_1$. Figure 7 shows the value of the function $f(1-q,\mu_1)$ for the example above as a function of its variables $1-q$ and μ_1. We can notice that for $\mu_1 = \overline{\mu}_1$ we get the linear random guess and for $\mu_1 = \underline{\mu}_1$ we get a steep curve reaching perfect classification.

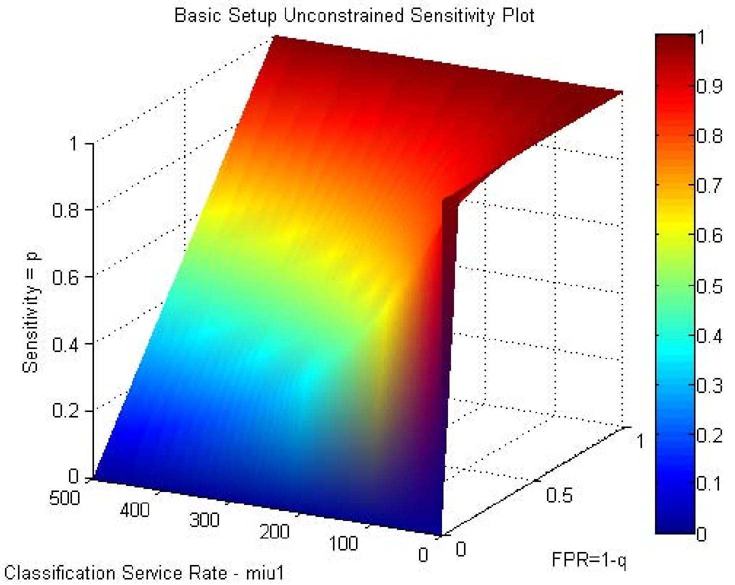

Figure 7. $f\left(1-q,\mu_1\right)$ as a function of its variables $1-q$ and μ_1 for the described scenario

Since the lower bound on the delay that is developed in the previous section, $(\overline{\mu_1}-\lambda_1)^{-1}+\mu_2^{-1}<\overline{W}$, does not depend on the values of λ_P and λ_N independently, but on the sum of the two, λ_1, which stays constant, all three scenarios have the same feasible range of $\overline{W}$ with respect to the input parameters, and it can be shown that it is: $\frac{1}{\overline{\mu}_1-\lambda_1}+\frac{1}{\mu_2}=0.0525<\overline{W}$.

Suppose that the acceptable delay assumes one of two levels, where the level is determined by the nature of the IR. *Tactical* IR, such as during tactical engagements on the battlefield or "ticking time bomb" scenarios has $\overline{W}=0.1$, which is close to the lower bound on the delay, while *strategic* IR, such as long term armament transactions, has $\overline{W}=5$, which means a delay of up to 5 time periods.

For the sake of brevity we name the six scenarios as follows: (i) tactical-low $\left(\overline{W}=0.1,\lambda_P=1\right)$, (ii) tactical-medium $\left(\overline{W}=0.1,\lambda_P=10\right)$, (iii) tactical-high

$\left(\overline{W} = 0.1, \lambda_P = 20\right)$, (iv) strategic-low $\left(\overline{W} = 5, \lambda_P = 1\right)$, (v) strategic-medium $\left(\overline{W} = 5, \lambda_P = 10\right)$ and (vi) strategic-high $\left(\overline{W} = 5, \lambda_P = 20\right)$.

3. Service Rate at the Analysis Station (μ_2)

The service rate, μ_2, at the analysis station is the bottleneck to which the classification adjusts its performance. For the discussed scenario, it can be shown that, given the value of $\overline{W}$, the feasibility conditions discussed in Chapter IV. A restrict the value of μ_2 to $\frac{1}{\overline{W} - \frac{1}{\overline{\mu}_1 - \lambda_1}} \approx 10.26 < \mu_2 < 110 = \lambda_1 + \frac{1}{\overline{W}}$ in the tactical scenario, and to $\frac{1}{\overline{W} - \frac{1}{\overline{\mu}_1 - \lambda_1}} \approx 0.2 < \mu_2 < 100.2 = \lambda_1 + \frac{1}{\overline{W}}$ in the strategic scenario.

The graph below (Figure 8) shows the effects of the value of μ_2 in its feasible range under the tactical IR for the different scenarios, and for the alternative of no classification (see (3.8)). As expected, the returns from an increased analysis rate diminish, and when the analysis rate reaches the upper bound, the sensitivity is nearly perfect since the analysis is no longer a bottleneck, thus almost all items are analyzed.

When costs are not taken into account, naturally the optimal sensitivity p^*, which is obtained from solving the model (3.10), is always better comparing the case of no classification, in which we denote the sensitivity by $p_{\min}$. Exploring the ratio $\frac{p^*}{p_{\min}}$, presented in Figure 9, allows us to evaluate the effect of classification for a given scenario.

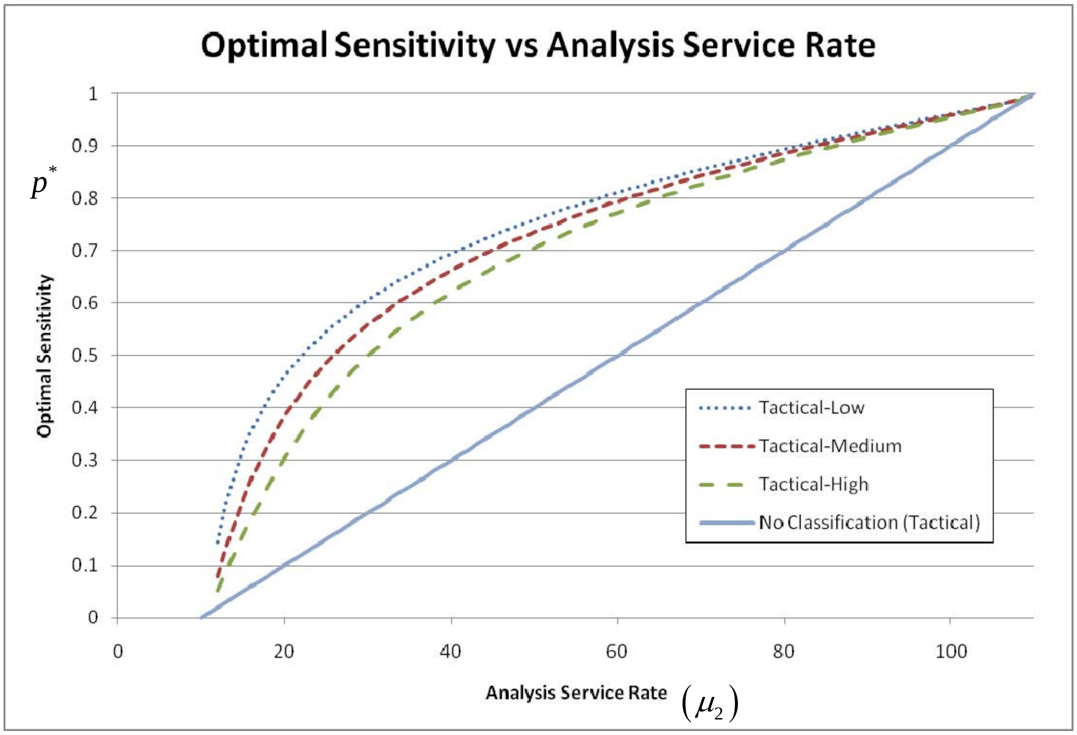

Figure 8. Optimal sensitivity vs. analysis service rate (μ_2)

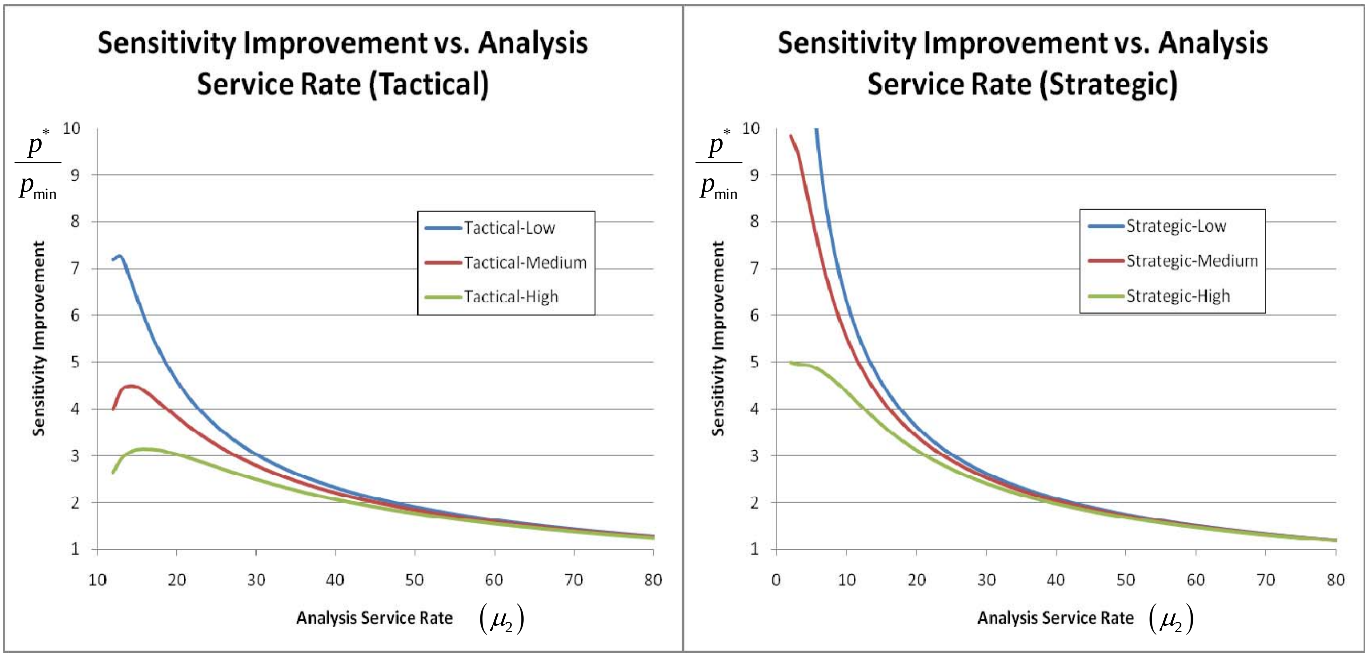

Figure 9. Sensitivity improvement vs. the analysis service rate (μ_2)
for the tactical (left) and strategic (right) scenario.

Operationally, this reinforces the intuition that classification is most effective when (1) the analysis poses a significant bottleneck, because its rate is significantly lower than the incoming rate of items, (2) the timeliness is not a crucial factor, and (3) the source has a low value in the sense that the rate $\frac{\lambda_P}{\lambda_1} \ll 1$.

4. Source Quality

The quality of the source is represented by the ratio $\frac{\lambda_P}{\lambda_1}$. To eliminate the effect of quantity, meaning the effect of an increase in λ_1 on the performance of the system, we iterate over the value of λ_P while keeping the sum λ_1 constant at 100. Recall that in the basic setup $\mu_2 = 20$ in any scenario, therefore if it was not for the classification, from (3.8) it would have followed that the minimal sensitivity is $p_{\min} = \frac{\mu_2 - \frac{1}{W}}{\lambda_1} = 0.1$ for the tactical IR and $p_{\min} = \frac{\mu_2 - \frac{1}{W}}{\lambda_1} = 0.198$ for the strategic IR, regardless of the value of λ_P. Figure 10 presents the sensitivity improvement $\frac{p^*}{p_{\min}}$ versus λ_P in the range $[1,100]$ for the *tactical* and *strategic* scenarios. Naturally, the higher the rate of positives is, λ_P, the lower the sensitivity improvement achieved by the classification.

The scenario in which the quality of a certain source varies over time is very reasonable in reality. When a source is first processed, it may bear large quantities of new information, namely positive, and at the extreme even $\lambda_P = \lambda_1$ is a possible scenario. However, the more this source is exploited, the less valuable it becomes due to repetitions of already known information. In the long run, this means that the source may stabilize on a certain rate of positives, which relates to the rate in which new information appears in that source.

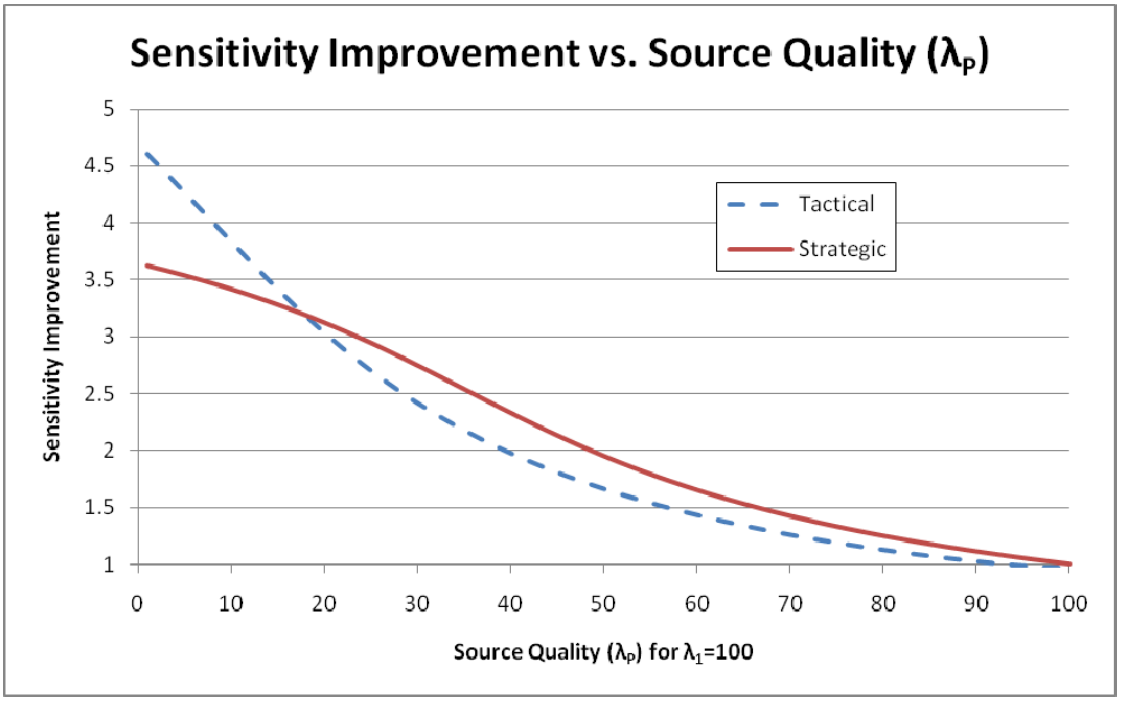

Figure 10. Sensitivity improvement factor versus source quality (λ_P where $\lambda_1 = 100$) under tactical and strategic IRs

The graph above reveals an interesting behavior: for low value sources, classification at the tactical scenario is more beneficial than in the strategic one. Nevertheless, around $\lambda_P = \mu_2$ the improvement curves cross and the classification in the strategic scenario is more beneficial than the one in the tactical.

Another insight can be drawn from the behavior of the decision variables $1-q$ and μ_1 at optimality. Recall that both represent the setting of the classification. Figure 11 shows the normalized classification rate, computed as $\frac{\mu_1}{\bar{\mu}_1}$ and the false positive rate $1-q$ for the different values of λ_P, when $\lambda_1 = 100$.

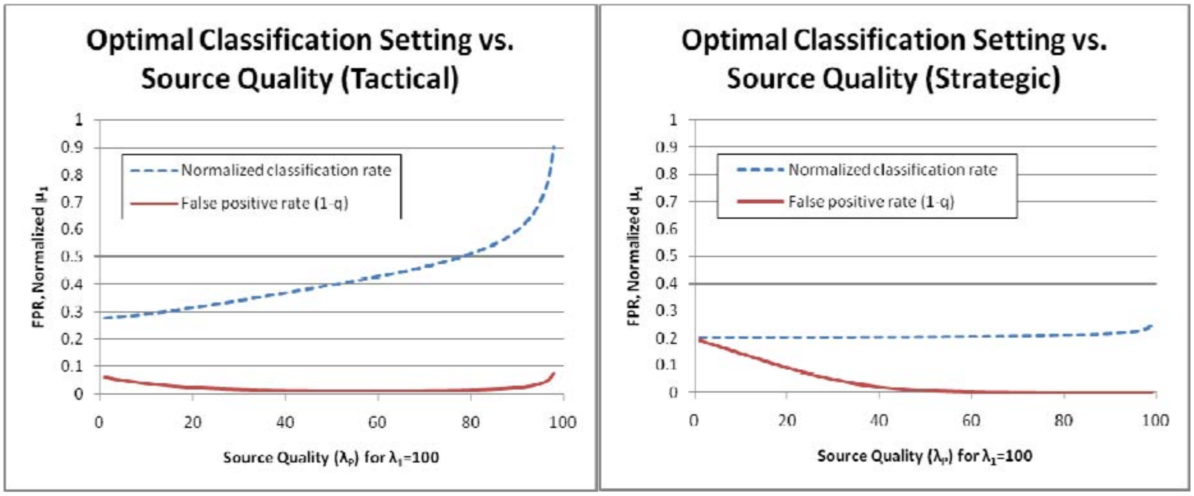

Figure 11. Classification setting (normalized classification rate, $\frac{\mu_1}{\bar{\mu}_1}$, and false positive rate, $1-q$, versus λ_P, where $\lambda_1 = 100$, under *tactical* IR (left) and *strategic* IR (right)

Comparing the graphs in Figure 11 reveals a different behavior at optimality between the two scenarios—tactical and strategic—as the source quality improves. In both scenarios, an increase in the quality of the source reduces the false positive rate at first, however, under the tactical scenario, the higher the quality of the source, the lower the quality is of the classification as manifested by the increase in the classification rate. On the other hand, under the strategic scenario, the quality of the classification is rather constant near its highest feasible value, while the threshold mechanism is relaxed so the false positive rate $1-q$ decreases. This difference shows the value of time under the different scenarios: while in the tactical scenario the optimal setting reduces the length of the test due to the urge to submit the item; in the strategic scenario the preference is to keep the test as accurate as possible, despite the required amount of time to do so.

5. Classification Capability Limit ($\underline{\varepsilon}$)

The classification capability limit, $\underline{\varepsilon}$, represents the ability of the classifier to address the questions correctly given an unlimited amount of time. This factor takes into account the experience and the training of the classifier: if the limit is at its maximum possible value then the classifier is as good as the analyst, and if it is at its minimum, the classifier is as good as a random guess.

Figure 12 reveals the impact of this parameter on the performance of the system as manifested by the degradation in the sensitivity when compared to a fully trained classifier, $\frac{p^*(\underline{\varepsilon})}{p^*(\underline{\varepsilon}=0)}$. For the same capability, the system performs better when classifying low value sources under strategic IR. Counter-intuitively, we get that the degradation is lower in the case where a high value source is classified.

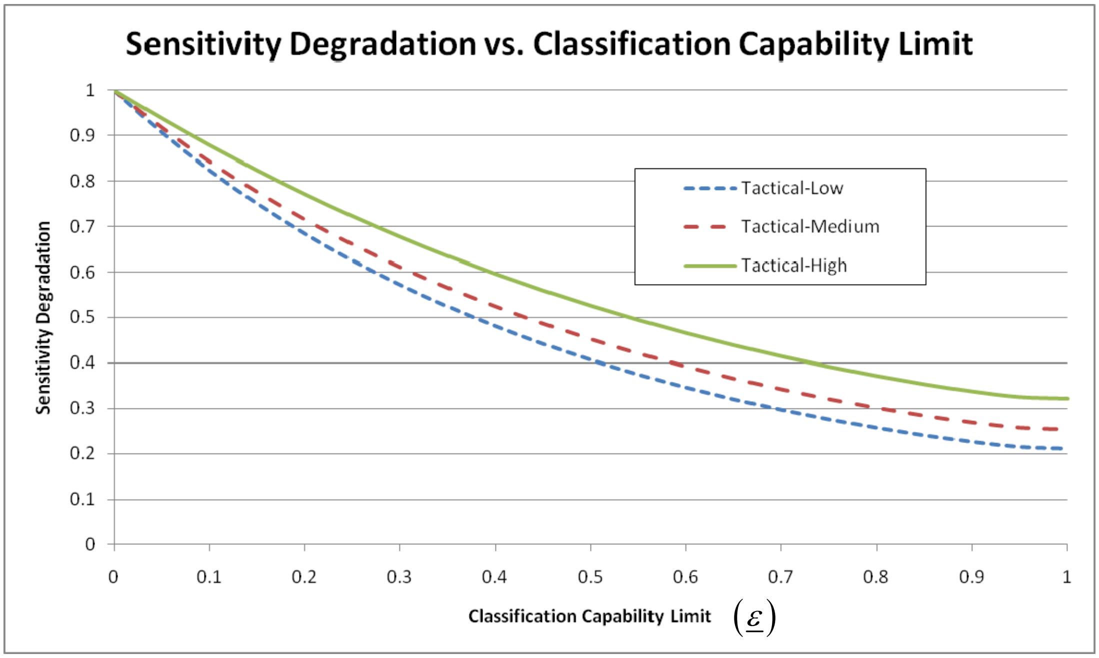

Figure 12. Sensitivity degradation vs. classification capability limit under *tactical* IR

C. COST-EFFECTIVENESS ANALYSIS

Figure 13 compares the expected cost per correctly identified positive item in a tactical scenario (see Chapter III Section E.1) for various values of costs *b* as a function of the source quality.

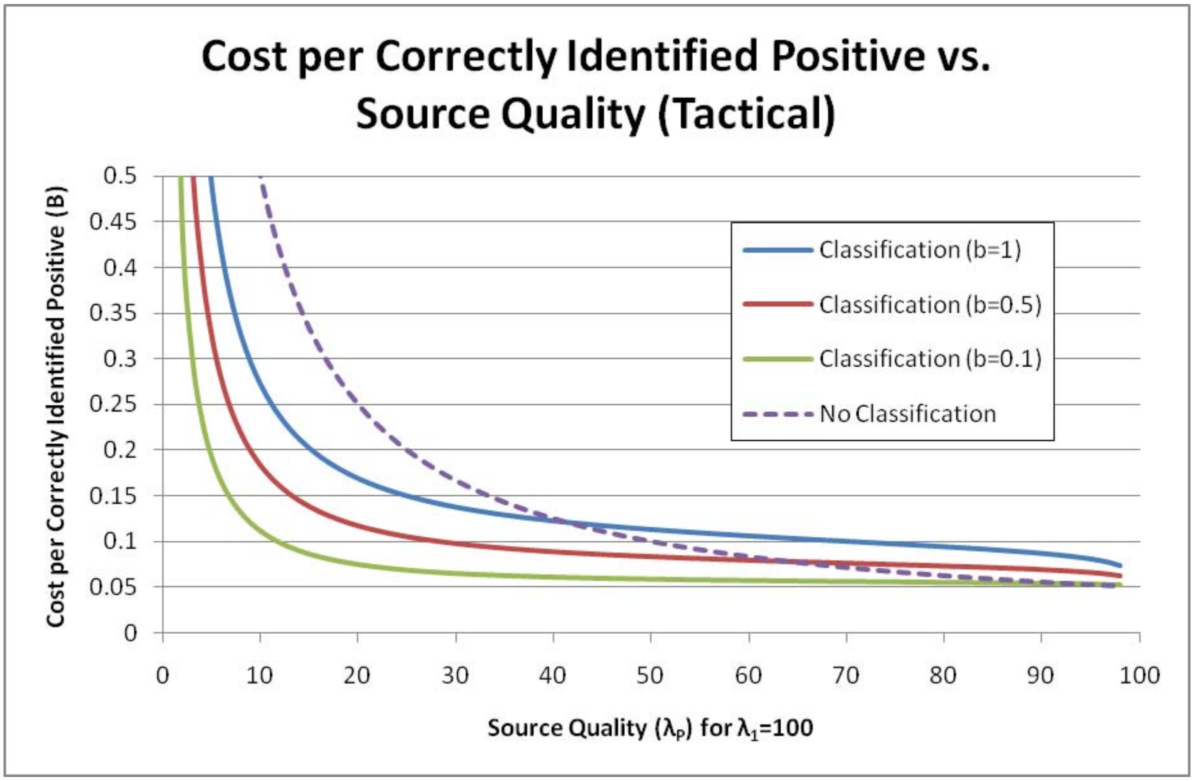

Figure 13. Cost per correctly identified positive (B) vs. the source quality under tactical IR, without classification and with classification with classifier cost $b = 0.1, 0.5, 1$.

Since $B \propto \frac{1}{\lambda_P}$, as the source quality λ_P increases, the cost per correctly identified positive decreases in all studied scenarios. However, the graph shows that the higher the quality of the source, the more economical it is to abandon the classification and allow analysts direct access to the items. As the quality of the source decreases, the implementation of a preliminary classification station outperforms in terms of cost-effectiveness, where the point of equality between the two options depends naturally on the cost of the classifier, *b*.

The operational meaning of this result is that when a source is first processed, it should be processed by the analysts until the rate of new information decreases below a certain level, which depends on the cost of the classifiers.

Since μ_2 and λ_P have an identical role in the expression for the cost per correctly identified positive item when no classification is implemented ($B = \frac{\lambda_1}{\mu_2 \lambda_P}$), and in the second term when classification is implemented ($B = \frac{b\lambda_1}{p\lambda_P\mu_1} + \frac{\lambda_2}{p\lambda_P\mu_2}$), similar behavior appears when we increase the analysis service rate, μ_2.

Another parameter which is of interest to explore in the context of cost-effectiveness is the classification capability limit, $\underline{\varepsilon}$. Assume that we are given several alternatives for the classifier position, each associated with their classification capability limit $\underline{\varepsilon}$ and cost *b*. Naturally the lower the classification capability limit, the higher the cost, until the point in which we are using an analyst equivalent classifier with $\underline{\varepsilon} = 0$ and $b = 1$. For a given scenario, such as the one described in the beginning of this section, we can make comparisons among the alternatives using their cost effectiveness or with respect to the alternative of no classification based on Figure 14.

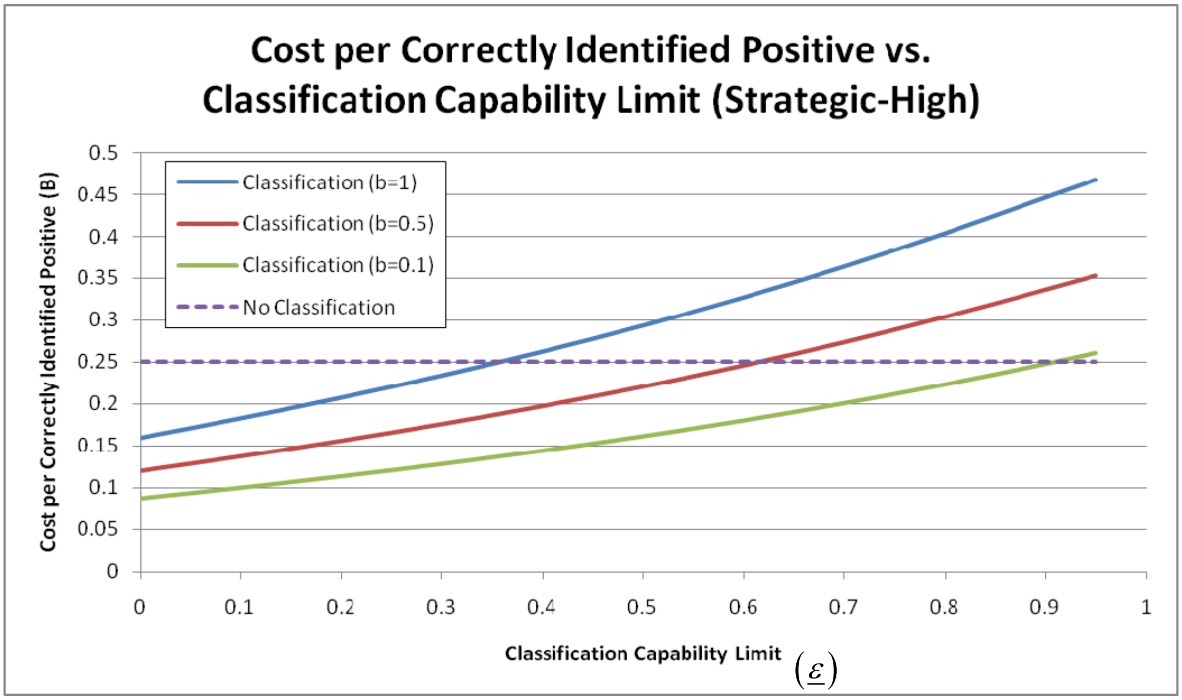

Figure 14. Cost per correctly identified positive vs. classification capability limit under a strategic IR with high value source.

Since the classification capability limit is a product of the training that the classifier receives and his accumulated experience, it can be determined to some extent when strategically planning the system. In Chapter V, we further discuss a possible extension to the classification model that will optimize this training under a given budget constraint.

V. CONCLUSIONS AND FUTURE WORK

A. CLASSIFICATION: SETTING AND MEASURES

The main contribution of this thesis is the formulation and analysis of a mathematical model for evaluating and optimizing the classification process in the intelligence cycle. The mathematical model, which is an optimization model based on a tandem queue, is used to determine tactical parameters in the classification process, while accounting for strategic parameters and studying the sensitivity of the optimal performance to those parameters. The key features are the ROC curve and its sensitivity to strategic parameters and the relations between the speed of the classification and its accuracy.

The model optimizes the sensitivity of the system using the classification setting parameters in two degrees of freedom: the length of time spent on each item's classification (service time) which defines the shape of the ROC curve of the classification process, and the implemented decision mechanism which determines the specific point on that ROC curve.

The effectiveness of the classification, as manifested by the sensitivity achieved at optimality, is measured with respect to the sensitivity of the system when no classification is implemented at all. This measure allows us not only to quantify the benefit of the classification under a given scenario, but also to compare the added value among different scenarios, allowing thumb rules for better classification resource allocation.

B. SENSITIVITY ANALYSIS

In the classification optimization model, the tactical parameters are adjusted to achieve optimality in terms of the sensitivity of the system. In the implementation of the model discussed in Chapter IV Section B, two scenario-related parameters are studied: the source quality λ_P and the expected delay constraint *W*. We have shown that the higher the source quality (with respect to a fixed arrival rate λ_1), the less beneficial it is to implement the classification. In addition, for low quality sources, better improvement is achieved for IRs that require immediate action; while for high quality sources the opposite applies: IRs that allow longer response time benefit more from the classification.

Another operational implication of the model is inferred from the different behavior at optimality between the tactical and strategic scenarios. As the source quality increases, in both scenarios the false positive rate should be minimized at first, in order to assure that the time share that analysts work on true positives is maximized. Nevertheless, at optimality, an increase in source quality results in a decrease in the optimal length of the test under a tactical scenario, while the test length stays rather constant in the strategic one.

Two main strategic parameters of the system are explored: the analysis service rate (μ_2) and the classification capability limit ($\underline{\varepsilon}$). The improvement achieved by implementing the optimal classification setting decreases as the analysis service rate increases and increases as the classification capability limit decreases. In addition, a cost-effectiveness study of the relationship between the classifier cost and the classification capability limit is developed as a means for comparing between different classifiers.

It is shown that when a source is highly valuable, it is more cost-effective to allow the analysts to directly interact with the source, despite the limited resources. Given the cost of the classification, the breakeven source quality in which both alternatives bear the same cost can be estimated.

C. FUTURE WORK

The model presented in this thesis forms a basic framework that can be extended in several directions that we believe are relevant and beneficial to the intelligence community. In this section, we discuss these directions briefly.

1. Partial Testing and Analysis

The model discussed in Chapter III assumed that a test, namely a set of questions to be addressed regarding a single item, must be fully completed before deeming it as positive or negative. A similar assumption is made concerning the bottleneck's processing, namely the analysis. This assumption is helpful when formulating a model that crystallizes the first order effects of the parameters. Nevertheless, a generalized model that allows different service time distributions for positive and negative items may result in a more precise description of reality and may be a step towards a predictive model.

2. System Optimization Models

The thesis' focus is to optimize the classification settings, namely the tactical parameters, in a given scenario and to compare the effectiveness among different scenarios. Although resources are not always interchangeable between the two phases – processing and analysis – due to organizational constraints, as discussed in Chapter II, in some cases resources may be shifted from one station to the other. In such a situation the need for system optimization models—models that optimize the strategic parameters in addition to the tactical ones—arises. We give here two variants of the system optimization model as examples for such models.

In the first example, the model is used to allocate specialists between the two phases given a constraint on the total number of specialists. In this case the model becomes a tandem M/M/k queue and the optimization is both on the tactical parameters and the number of servers in each station.

In the second model, the cost per time unit of a classifier is functionally related to the classification capability limit, where higher costs are associated with a lower classification capability limit, in the sense that perfect classification can be achieved. In this model, an additional constraint is added to limit the cost-effectiveness of the system. The explored tradeoff in this case would be between the classifier's quality and the time spent on classifying each item.

3. Bias: Deception and Misconception

A discussion about intelligence cannot be complete without discussing the effects of bias that can emerge from deception that is the result of an active counter-intelligence effort, or misconception, that is, erroneously adopted beliefs by senior intelligence experts and decision makers. In the presented model, the notion of positives and negatives is defined by the analysts’ perception and instructions. Such an assumption aligns with user satisfaction measures, since the analysts are the processing phase’s user, whose satisfaction is measured by the sensitivity of the classification to the given instructions, regardless of their absolute value.

Nevertheless, it may be of interest to incorporate the bias as a parameter of the model, and quantify its effects on the absolute value of the intelligence production process, as viewed by the opponent. This sort of model requires a model for the process in which the analyst uncovers the true nature of each item by a process of exploration. At any given time, the analyst has to divide his time between exploration and exploitation. Exploration means that items are processed even if they are declared as negatives, in order to find clues for unknown positives. On the other hand, when exploiting, only items deemed as positives are processed.

LIST OF REFERENCES

Bose, R. (2008). Competitive intelligence process and tools for intelligence analysis. *Industrial Management & Data Systems*, 108(4), 510–528.

Brooke, A., Kendrick, D., Meerus, A., Raman, R., & Rosenthal, R. E. (1998). *GAMS User's Manual*. © GAMS Development Corporation.

Card S., & Pirolli P. (2005). The Sensemaking Process and Leverage Points for Analyst Technology as Identified through Cognitive Task Analysis. *Proc. International Conference on Intelligence Analysis*.

Coffman T., Greenblatt S., & Marcus S. (March 2004). Graph-based technologies for intelligence analysis. *Communications of the ACM, 47(3)*, 45–47.

Davis, J., & Goadrich M. (2006). The relationship between precision-recall and ROC curves. *Proceedings of the 23rd international conference on machine learning (ICML)*.

Defense Science Board (2009). *Report of the Defense Science Board Advisory Group on Defense Intelligence: Operations Research Applications for Intelligence, Surveillance and Reconnaissance (ISR)*. Washington, D.C.: Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics, January 2009. http://www.acq.osd.mil/dsb/reports/ADA493773.pdf.

Drud, A. (2005). *CONOPT*. ARKI Consulting and Development A/S. Bagsvaerd, Denmark. http://www.gams.com/solvers/conopt.pdf.

Fawcett T. (June 2006). An introduction to ROC analysis. *Pattern Recognition Letters - Special issue: ROC analysis in pattern recognition archive, 27(8)*, 861–874.

Greitzer F. (2005). *Methodology, Metrics and Measures for Testing and Evaluation of Intelligence Analysis Tools*. Tech report PNWD-3550, Battelle-Pacific Northwest Division.

Heuer R. J. (2001). *Psychology of Intelligence Analysis*. Center for the Study of Intelligence, Central Intelligence Agency.

Horowitz B. M., & Haimes Y. Y. (2003). Risk-based methodology for scenario tracking, intelligence gathering, and analysis for countering terrorism. *Systems Engineering, 6*, 152–169.

Hulnick A. S. (2006). What's wrong with the Intelligence Cycle. *Intelligence and National Security, 21(6)*, 959–979.

Järvelin K., & Kekäläinen J. (October 2002). Cumulated Gain-Based Evaluation of IR Techniques. *ACM Trans. Information Systems*, Oct. 2002, 422–446.

Johnson L. K., & Wirtz J. J. (2004). *Strategic Intelligence: Windows Into a Secret World*. Los Angeles, CA: Roxbury Publishing Company.

Johnston R. (2005). Analytic Culture in the US Intelligence Community: An Ethnographic Study. *Center for the Study of Intelligence, Central Intelligence Agency*. https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/books-and-monographs/analytic-culture-in-the-u-s-intelligence-community/full_title_page.htm

Kahaner L. (1998). *Competitive Intelligence: How to Gather, Analyze and Use Information to Move your Business to the Top*. Touchstone, New York, NY.

Kaplan E. H. (2010a). Terror queues. *Operations Research*, in press (doi:10.1287/opre.1100.0831)

Kaplan E. H. (2010b). Intelligence operations research. *Operations Research*, in press.

Marzban C. (2004). The ROC Curve and the Area under It as Performance Measures. *Weather and Forecasting, 19(6),* 1106–1114.

Miller J. O., Pawling C. R., & Chambal S. P. (2004). Modeling the U.S. Military Intelligence Process, *Defense Technical Information Center,* http://handle.dtic.mil/100.2/ADA466314

Paté-Cornell M. E. (2002). Fusion of Intelligence Information: A Bayesian Approach. *Risk Analysis, 22(3),* 445–454.

Pinker E. J., Szmerekovsky J. G. & Tilson V. (July 2009). Managing a Secret Project. Simon School Working Paper No. FR 09-17. Available at SSRN: http://ssrn.com/abstract=1434696

Pirolli P., & Card S. K. (1999). Information foraging. *Psychological Review, 106*, 643–675.

Pirolli P. (2009). Powers of 10: Modeling complex information-seeking systems at multiple scales. *IEEE Computer, 42*, 33–40.

Provost F., Fawcett T., & Kohavi R. (1998). The case against accuracy estimation for comparing induction algorithms. *Proceeding of the 15th International Conference on Machine Learning,* 445–453.

Richelson J. (1999). *The U.S. Intelligence Community*. Cambridge, MA: Ballinger Publishing Company. Second Edition.

Russell D. M., Stefik M. J., Pirolli P., & Card S. K. (1993). The cost structure of sense-making. Paper presented at the INTERCHI '93 *Conference on Human Factors in Computing Systems*, Amsterdam.

Singhal A. (2001). Modern Information Retrieval: A Brief Overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 24(4),* 35-43. http://singhal.info/ieee2001.pdf

Skroch E. (2005). *Interdicting a nuclear weapons project*. Master's thesis, Operations Research Department, Naval Postgraduate School, Monterey, CA.

Steele J. M. (1989). Models for managing secrets. *Management Science, 35,* 240–248.

Swets J. (1988). Measuring the accuracy of diagnostic systems, *Science, 240,* 1285–1293.

Yue W. T., & Cakanyildirim M. (December 2010). A cost-based analysis of intrusion detection system configuration under active or passive response. *Decision Support Systems*, *50(1),* 21–31.

THIS PAGE INTENTIONALLY LEFT BLANK

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California

3. Professor Moshe Kress
   Department of Operations Research
   Naval Postgraduate School
   Monterey, California

4. Professor Roberto Szechtman
   Department of Operations Research
   Naval Postgraduate School
   Monterey, California

5. Professor Patricia Jacobs
   Department of Operations Research
   Naval Postgraduate School
   Monterey, California

6. Professor Edward H. Kaplan
   Yale School of Management
   Yale University
   New Haven, Connecticut

7. Yeo Tat Soon
   Director, Temasek Defence Systems Institute (TDSI)
   National University of Singapore
   Singapore

8. Ms Tan Lai Poh
   Temasek Defence Systems Institute (TDSI)
   National University of Singapore
   Singapore